Estimating the change in meta-analytic effect size estimates after the application of publication bias adjustment methods

Article (Accepted Version)

Estimating the change in meta-analytic effect size estimates after the application of publication bias adjustment methods

Martina Sladekova[1], Lois E. A. Webb[1], and & Andy P. Field[1]

[1] School of Psychology

University of Sussex

Author Note

Martina Sladekova. ORCiD: https://orcid.org/0000-0001-5059-6576

Andy P. Field. ORCiD: https://orcid.org/0000-0003-3306-4695

Study design and the analysis plan were preregistered prior to accessing the data: https://osf.io/kxjs3 . Data and materials are available at https://osf.io/k9hqm/ (non-editable time-stamped version of the repository can be found at https://osf.io/kmdr6). Publications associated with the raw datasets analysed in this study were previously reviewed for reporting practices by Robert A. T. Avery and Maram Kamar, Martina Sladekova and Andy P. Field: https://osf.io/ruvhd/ . The analyses run in these two studies do not reuse the same data. The data appearing in the manuscript were uploaded into the OSF repository prior to submission for publication.

The authors made the following contributions. Martina Sladekova: Conceptualisation, data extraction, data reanalysis, main data analysis, original manuscript preparation, review and editing; Lois E. A. Webb: Data extraction and reanalysis; Andy P. Field: Conceptualisation, manuscript review and editing, project supervision.

Correspondence concerning this article should be addressed to Martina Sladekova, School of Psychology, University of Sussex, Brighton, UK. E-mail: m.sladekova@sussex.ac.uk

Abstract

Publication bias poses a challenge for accurately synthesising research findings using meta-analysis. A number of statistical methods have been developed to combat this problem by adjusting the meta-analytic estimates. Previous studies tended to apply these methods without regard to optimal conditions for each method's performance. The present study sought to estimate the typical effect size attenuation of these methods when they are applied to real meta-analytic datasets that match the conditions under which each method is known to remain relatively unbiased (such as sample size, level of heterogeneity, population effect size, and the level of publication bias). 433 datasets from 90 papers published in psychology journals were reanalysed using a selection of publication bias adjustment methods. The downward adjustment found in our sample was minimal, with greatest identified attenuation of $b = –0.032$, 95% Highest Posterior Density interval (HPD) ranging from –0.055 to –0.009, for the Precision Effect Test (PET). Some methods tended to adjust upwards, and this was especially true for datasets with a sample size smaller than ten. We propose that researchers should seek to explore the full range of plausible estimates for the effects they are studying and note that these methods may not be able to combat bias in small samples (with less than ten primary studies). We argue that although the effect size attenuation we found tended to be minimal, this should not be taken as an indication of low levels of publication bias in psychology. We discuss the findings with reference to new developments in Bayesian methods for publication bias adjustment, and the recent methodological reforms in psychology.

Keywords: Meta-analysis, Publication bias, Bias adjustment, Effect size attenuation

Word count: 10581

Estimating the change in meta-analytic effect size estimates after the application of publication bias adjustment methods

The use of meta-analysis has grown exponentially in the past two decades (Ioannidis, 2016). It allows researchers to statistically synthesise findings, identify common moderators of effects, plan future studies, and inform policy change (Hunter & Schmidt, 1996; Maki et al., 2018). Like any statistical method, meta-analysis can be affected by various sources of bias, resulting in an over- or underestimation of the effect in question. Bias is commonly introduced into the analysis by, for example, failing to account for the dependence of the effect sizes in the meta-analysis (Cheung, 2014), inadequately correcting measurement artefacts (Hunter & Schmidt, 2004), or neglecting the statistical assumptions of the applied models (Kontopantelis & Reeves, 2012). A sample can also become biased when the acquired primary studies are not representative of the full range of results produced within a research field. This can be related to an inadequate sampling strategy, but also to bias in the publication process (Ferguson & Heene, 2012).

Publication Bias

Publication bias, or the "file drawer problem" (Rosenthal, 1979) occurs when the probability of a study being published favours one type of result over another, distorting the scientific record. Null Hypothesis Significance Testing (NHST) is a dominant hypothesis testing approach in psychology, and the binary labelling of p values as statistically significant or non-significant is a defining element of publication bias. Studies which confirm the researchers' predictions and produce p values below the conventional threshold of .05 are more likely to get published than studies producing statistically non-significant results where $p > .05$. (Fanelli, 2010; Sterling et al., 1995). This censoring can then happen on two levels. On the journal level, a study may be rejected by the editors or the reviewers because non-significant results are deemed uninteresting or difficult to interpret (Greenwald, 1975;

Sterling et al., 1995). On an individual level, a researcher may decide not to submit a study for publication, either because the results do not support their theory, or because they believe that a statistically non-significant result would not be accepted for publication. Either way, the study ends up in the researcher's file drawer and its inclusion in subsequent meta-analyses depends on the meta-analysts' access to these records and the researchers' willingness to share their data (Ferguson & Heene, 2012). Considering that small effects are less likely to cross the threshold of statistical significance if a study is inadequately powered (Cohen, 2013; J. Cohen, 1992), the meta-analysed effect may then show an upward bias (when considering the absolute value of the effect size) because large effects get published, while small effects do not. This highlights another problem emerging from publication bias - the small study effect (Sterne et al., 2000). Studies with small sample sizes tend to produce inflated effects and increase the upward bias of the meta-analytic estimate. While common meta-analytic methods address this problem to an extent by assigning smaller weights to studies with greater variance Borenstein et al. (2011), the effect remains problematic, especially if small-sample studies comprise a large proportion of the meta-analytic sample (Sterne et al., 2000).

The extent of publication bias in psychology is unknown, however an indication comes from the investigations of the statistical power. Power is the probability of detecting an effect of a given magnitude as statistically significant at a predetermined alpha level, assuming that the alternative hypothesis is true (Cohen, 2013). The average power in psychology is approximately 50% (Cohen, 2013; J. Cohen, 1992; Cumming, 2013), meaning that out of n replications, half will not detect the investigated effect as statistically significant. Yet, if we look at the published record, we find that over 90% of studies report statistically significant findings (Fanelli, 2010; Scheel et al., 2020), which is at least 40 percentage points more than we should expect.

Methods for Addressing Publication Bias

Several statistical methods for detecting the presence of publication bias in the sample exist. The Fail-Safe Number (FSN, Rosenthal, 1979) estimates the number of unpublished studies that would need to be included in the meta-analysis for the result to change to a statistically non-significant one. Similarly, Orwin's FSN estimates the number of studies needed for the effect size to drop below a specified value (Orwin, 1983). Other detection tests include the Egger Regression Test (Egger et al., 1997), or the Rank-Correlation Test (Begg & Mazumdar, 1994), both of which are statistical tests assessing the symmetry of the funnel plot (Figure 1, Egger et al., 1997), or the Test of Excess Significance, which computes whether the observed number of statistically significant studies is different from the expected count (Ioannidis & Trikalinos, 2007).

Recently, the focus has switched to methods that adjust the meta-analytic estimate for the presence of publication bias, rather than try to detect its presence in the sample. This approach might be preferable for two reasons. First, like all statistical methods, detection methods that rely on NHST might be overly sensitive in large samples and not sensitive enough in small samples (Field, 2018; Zimmerman, 2004)[1]. Second, considering the over-representation of statistically significant results in the literature with reference to the average statistical power in psychology, publication bias is likely to be present, unless in the context of meta-analyses of large-scale multi-lab sets of replications (Carter et al., 2019), or meta-analyses dealing with registered reports (Nosek & Lakens, 2014; Scheel et al., 2020).

Methods investigated in the present study include Trim-and-Fill (TF, Duval, 2005), Precision-Effect Test (PET, Stanley, 2008), Precision-Effect Estimate with Standard Error (PEESE, Stanley, 2008), PET-PEESE (Stanley, 2017), Weighted Average of the Adequately Powered studies (WAAP-WLS, Stanley et al., 2018) and selection models (McShane et al.,

[1] Although the FSNs (Orwin, 1983; Rosenthal, 1979) do not perform significance testing, the methods have been criticised for making unrealistic assumptions about unpublished studies and returning large estimates, providing the researchers with a false sense of confidence (Ferguson & Heene, 2012).

2016; Vevea & Woods, 2005).

Trim-and-Fill

TF utilises the funnel plot (Figure 1), which plots the primary effect on the x axis and a measure of sampling variation on the y axis. Doing so should result in a symmetrical funnel shape (Figure 1a) - unless publication bias is present, in which case the distribution of the data points will be asymmetrical (Figure 1b). TF uses an imputation method to estimate which unpublished effects are missing. First, the number of missing studies k_0 is estimated based on the symmetry of the distribution of the effects around the unadjusted estimate. Second, k_0 studies are trimmed off the right half of the plot to make it more symmetrical, and the effect is re-estimated based on the trimmed data. Using this new central estimate, k_0 is re-estimated and the trimming process is repeated until the estimate stabilises. Finally, k_0 missing studies are imputed to the left half of the plot symmetrically around the effect size estimate from the latest iteration, and the final adjusted effect is re-estimated (Duval, 2005). Funnel plot and TF have been repeatedly criticised as tools for addressing publication bias, as funnel plot asymmetry can also be the result of missing moderators and heterogeneity among the effects, especially when the primary effect sizes are correlated with their standard errors (Ioannidis, 2008; Lau et al., 2006; Peters et al., 2010; Sterne et al., 2011). In spite of that, the method remains prevalently used.

PET, PEESE, and PET-PEESE

PET, PEESE, and PET-PEESE work on the principle of including the standard error or its squared value as a predictor in the meta-analytic model. In a standard meta-analysis, the effect size is predicted only from the weighted intercept. When PET is applied, inverse variances of the primary studies are used as weights, while the standard error of primary effect sizes is added as a predictor. In an equation form, this means:

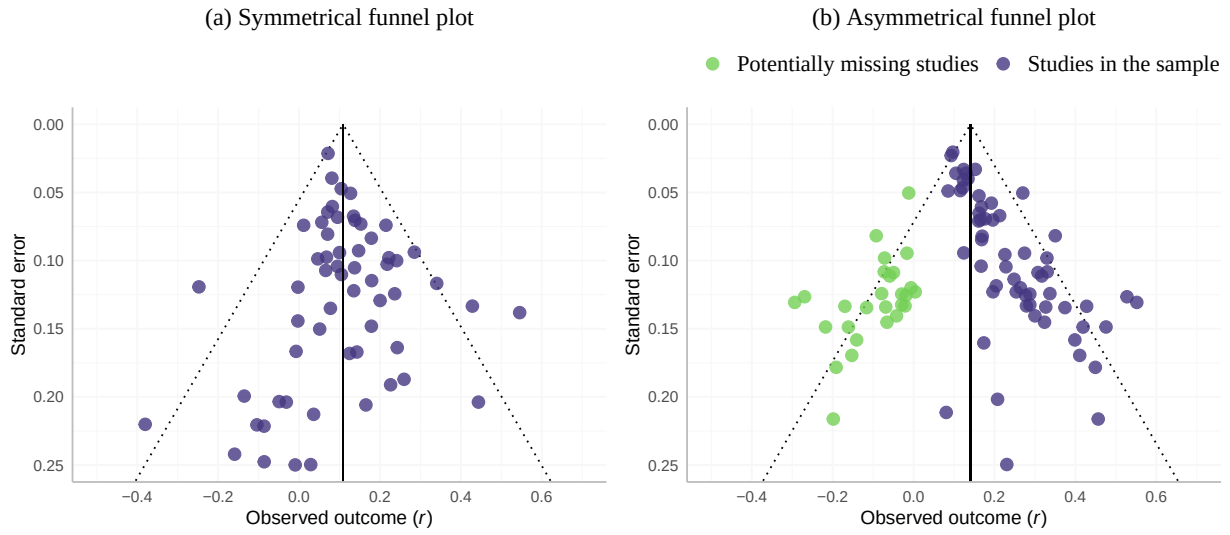

Figure 1

Example of (a) a symmetrical funnel plot and (b) an asymmetrical funnel plot. With little or no publication bias, the plot is expected to be symmetrical (although see Sterne et al. (2011) for a discussion on the interpretation of funnel plot asymmetry with reference to heterogeneity of the effects and missing moderators). Funnel plot data were obtained from https://osf.io/rf3ys/ (Carter et al., 2019).

$$\hat{d}_i = \hat{b_0} + \hat{b_1} se_i + e_i$$

Once the model is fitted, the coefficient value of the intercept is taken as the meta-analytic estimate. PEESE also uses the inverse variance as weights, however the modelled relationship between the effect size and its standard error is quadratic:

$$\hat{d}_i = \hat{b_0} + \hat{b_1} se_i^2 + e_i$$

PET-PEESE was developed to balance the bias identified in the performance of PET and PEESE. When the underlying population effect is zero, PET tends to outperform

PEESE, whereas when the population effect is different form zero, PEESE shows better performance (Stanley, 2017). PET-PEESE evaluates the statistical significance of the PET estimate to determine which estimate to use as final. If PET is non-significant (i.e. the intercept is not significantly different from zero), the PET estimate is taken. If PET is significant, PEESE estimate is taken as the adjusted estimate.

WAAP-WLS

As a first step, WAAP-WLS obtains an estimate from an intercept-only weighted-least-squares (WLS) model. In the second step, the power to detect the first estimate is retrospectively calculated for each primary study. A follow-up meta-analysis is then performed on studies that have at least 80% power to detect the original WLS estimate. The resulting estimate is therefore the weighted average of adequately powered studies (WAAP). If the sample contains fewer than two adequately powered studies, the first WLS estimate is used as a fall-back option (Stanley et al., 2018).

Selection Models (3PSM and 4PSM)

Also known as the Weight Function Models (Vevea & Woods, 2005), selection models consider 3 parameters (3PSM) or 4 parameters (4PSM) as part of the model when computing the adjusted estimate. Parameters in 3PSM include the primary effect size, which provides the population estimate; a heterogeneity parameter, and a weight parameter. The weights are assigned based on the assumption that a p value associated with a primary study can affect its chances of publication and subsequent inclusion in a meta-analysis. The weights are modelled as a step function with a single cut-point at $p = .05$ for a two-tailed test. This splits the range of possible p values into two bins - one with statistically significant p values and one with non-significant p values , wherein the weights for the studies in "non-significant" bin are assigned greater weights to account for the unpublished studies missing from the meta-analytic sample. For 4PSM, two cut-points for p values are

defined - the additional bin created includes p values in the "marginally significant" range between 0.05 and 0.10. This represents the possibility that studies with p values in this range have higher chance of being published than studies with greater p values. The parameters in the model are estimated with maximum likelihood (McShane et al., 2016).

p-*curve and* p-*uniform*

p-curve and p-uniform consider the frequency distribution of all statistically significant p values from the meta-analytic sample (Simonsohn et al., 2014). When the null hypothesis for a given effect is true (i.e. when the effect under study is zero), this frequency distribution is expected to be uniform between 0 and .05 as the studies are assumed to have no statistical power. As the statistical power increases, the distribution, or the "p-curve" becomes increasingly right-skewed. Only true effects are therefore expected to generate a right-skewed p-curve, as these will contain greater frequency of p values in the area of $p = 0.01$, whereas studies with larger p values closer to the conventional threshold of 0.05 would be expected to occur less frequently. Conversely, if p values around 0.05 occur more frequently and lessen the degree of right skew, this is assumed to be indicative of some degree of publication bias and therefore a possible overestimation of the meta-analytic effect. p-curve and p-uniform use the degree of skewness to adjust the effect size. The effect size estimate that yields the smallest distance between the observed distribution of p values and the uniform distribution is taken as the final estimate. For the p-curve method, this distance is estimated using the Kolmogorov-Smirnov statistic, whereas the p-uniform method applies an estimator based on the Irwin-Hall distribution (Van Assen et al., 2015). As the two methods differ only in the estimation algorithm, they are expected to show similar performance (McShane et al., 2016; Van Aert & Van Assen, 2018).

Assessing Method Performance

The methods outlined above routinely outperform the standard unadjusted random-effects or fixed-effects meta-analytic models in terms of precision of the estimated effect sizes (Table 1). However, each method may show suboptimal performance under a specific set of circumstances, and the adjusted estimates may still over- or underestimate the true effect size. Table 1 contains an overview of the adjustment methods and the summary of the conditions under which the methods underperform, identified in simulation studies. Overall, heterogeneity is problematic for all the methods to a varying extent. For methods like PET-PEESE and WAAP-WLS, small sample sizes can result in a biased estimate. This is because PET-PEESE relies on statistical significance testing and as such its power to detect a potentially problematic relationship between the effect size and the standard error can be low in small samples. WAAP-WLS on the other hand requires a number of adequately powered primary studies in the sample in order to provide an adjusted estimate. The smaller the sample, the smaller the chance that it includes studies with adequate power. What constitutes a "small sample" can differ from one study to another. In prior simulation studies summarised in Table 1, this often constitutes samples with fewer than ten primary studies. It is worth noting however, that for PET-PEESE this will also depend on the strength of the association between the effect size and the standard error, whereas for WAAP-WLS this will depend on the number of primary studies with adequate power (for example a sample of ten with eight well-powered primary studies in the sample would provide a more accurate estimate than a sample of ten with only two primary studies with adequate power). The bias caused by heterogeneity and small samples tends to be upward, and the extent differs across the methods.

The ways of simulating publication bias censoring functions and sample parameters vary. Because of this, comparing the performance of these methods relative to each other across simulation studies can be challenging. Carter et al. (2019) brought the methods discussed above into a single simulation that evaluated the performance of each method

under a range of conditions typically found in psychology research. Carter et al. (2019) concluded that no single method can be recommended as the best performing when compared to the other methods. Instead, the authors recommend that for each individual meta-analysis, the adjustment method should be selected depending on the values of the parameters unique to the meta-analytic dataset at hand, like the sample size, heterogeneity, assumed level of publication bias or assumed population effect size.

Research Objectives

Simulation studies are able to evaluate bias in the performance of the adjustment methods with reference to a true population effect. Simulations, however, provide don't provide information about the kind of change in the effect size that should be expected if the methods were applied across the field in accordance with recomendations, or what kind of change in the estimates a meta-analyst should expect when they apply the method to their own data. There are two main reasons for this uncertainty. Although a simulation can be informative in guiding the process of selecting the optimal method, the procedure works with simplified values and distributions that cannot fully capture the complexity of the sample level or the population level characteristics across the field. More importantly, each adjustment method evaluated in a simulation will always perform with a certain level of bias, even if a method is evaluated as best performing out of all the methods under consideration. This bias is meaningful only with reference to a known population value. In real research context, the population parameter value is unknown and it is therefore impossible to tell what level of adjustment should be expected as a result of applying a given method. This uncertainty is exacerbated by the fact that the bias for a single method will vary across unknown population level characteristics, and that the bias will also differ across the the different adjustment methods. Re-analysis of exisiting published meta-analyses and the application of these methods to real datasets can therefore contribute unique insights into how the landscape of published findings could change under scenarios where publication bias

is attended to in accordance with exisitng guidance. Van Aert et al. (2019) applied a selection of these methods to a subset of homogeneous meta-analytic samples from medical and psychological studies. They reported minimal change in the estimates after the application of the alternative methods. However, homogeneous samples are unlikely to be representative of the samples commonly found in psychology (Van Erp et al., 2017), and the application of these methods did not take into consideration the sample level and the population level variables which can impact the method performance. The present study builds on the findings from Carter et al.'s (2019) simulation and selectively applies the appropriate publication bias adjustment methods to a sample of 433 meta-analytic datasets. The key objectives of this study are to estimate (1) the typical attenuation in the published meta-analytic effects after the application of the methods adjusting for the presence of publication bias, (2) the plausible range of this attenuation, and (3) the plausible limits of the variation in the effect size attenuation across the different adjustment methods and scenarios typical for psychology research.

Table 1

Summary of simulation studies evaluating the performance of publication bias adjustment methods

Method	Description	Performance
Trim-and-fill	Iteratively estimates the number of missing studies based on the symmetry of the funnel plot. Imputes the missing studies and re-estimates the effect size.	**van Assen et al. (2015)**: slight underestimation of the effect size when no publication bias is present, and overestimation under strong publication bias; upward bias under heterogeneity. **Moreno et al. (2009)**: suboptimal performance under heterogeneity **Carter et al. (2019)**: increased overestimation under strong publication bias, increased overestimation under heterogeneity.
PET, PEESE & PET-PEESE	Fits a linear (PET) or a quadratic (PEESE) relationship between the effect size and the standard error. PET-PEESE uses the statistical significance of the PET estimate to decide which estimate to use. If PET is statistically significant, PEESE is used, and vice-versa.	**Stanley (2017)**: Bias in small sample sizes ($k < 20$) and when heterogeneity is high. **Carter et al. (2019)**: Unbiased under strong publication bias and with small sample size ($k = 10$), so long as heterogeneity is low; small downward bias under high heterogeneity.
WAAP-WLS	A WLS meta-analysis is performed on studies that have at least 80% statistical power to detect the WLS estimate obtained when all studies are included. The latter is used as a fallback option if there are fewer than two adequately powered studies.	**Stanley & Doucouliagos (2016)**: WLS outperforms random-effects models when publication bias is present. **Stanley (2017)**: Bias when sample size is small, reduced bias with increasing sample size; increased bias under heterogeneity. **Carter et al. (2019)**: Very slight overestimation under no heterogeneity, increase in upward bias after adding heterogeneity.
Selection models (3PSM & 4PSM)	Model includes the effect size parameter, the heterogeneity parameter and the weight parameter. Weight parameter is modelled as a step function which assigns the likelihood that a non-significant study gets published. 3PSM divides p values into two bins, 4PSM divides them into three bins.	**McShane et al. (2016)**: Slight underestimation for null and small effects, and when the sample size is small. Relatively unbiased under heterogeneity. **Carter et al. (2019)**: Unbiased under no heterogeneity and strong publication bias. Also remains unbiased under heterogeneity.
p-curve & p-uniform	Plots the distribution of statistically significant p-values in the meta-analytic sample. Uses the degree of right skew to test the null hypothesis and estimate the adjusted effect.	**Simonsohn et al (2014)**: Accuracy does not depend on k or heterogeneity. **van Aert & van Assen (2018)**: Slight downward bias for smal k. **van Assen et al. (2015)**: Upward bias under heterogeneity. **Carter et al. (2019)**: Substantial upward bias under heterogeneity.

Methods

Sample

The project was conducted in two phases. In phase one, a random sample of meta-analyses published in years 2008 and 2018 was selected[2]. A range of reporting practices present in the selected papers was recorded using a coding scheme developed for the purposes of this project. Details of the coding scheme, analyses run on the coded dataset, and inclusion/exclusion criteria can be found at https://osf.io/ruvhd. In total, reporting practices of 169 papers were coded in phase one of the project. This paper reports the results of phase two, which involved extraction of raw data from the papers coded in phase one. In addition, authors of included papers were contacted with requests for data. Obtained samples were then reanalysed with publication bias adjustment methods.

Inclusion Criteria

A study was included in the sample if (i) it was included in phase one of the project, (ii) raw data were extractable from tables or figures, supplemental materials, or were made available by the authors, (iii) primary effect sizes were reported as correlation coefficients r or the data contained information needed to transform primary effect sizes into r, and (iv) data included variance of the primary studies or provided information necessary for variance estimation.

---

[2] The present study was a follow-up from a review conducted in the first phase by Avery et al. (2020) (manuscript in preparation) who looked at the change in meta-analytic reporting practices for the two years included in the sample. Data related to the first phase, including the details about the sampling process are publicly available on the OSF (https://osf.io/ruvhd/).

Exclusion Criteria

In addition to failing to meet the above inclusion criteria, studies were excluded if (i) the study was a meta-meta-analysis or a meta-analysis using internal databases as opposed to research papers, as the character of publication bias in these types of samples is unclear, or (ii) the original analyses failed to reproduce.

Figure 2 displays the selection process for the studies included in the final sample. 118 researchers were contacted with requests for data, of whom 27 were able to provide datasets, while 60 no longer had access to the data, and the remaining 31 did not respond. Of all the papers, 78 reported compatible primary effect sizes in tables, figures or supplemental materials, resulting in 105 papers with extractable information, containing 465 separate datasets. 32 datasets were excluded because the results either did not reproduce (n = 22), effect sizes could not be converted into the correlation coefficient ($n = 9$), or the study used incompatible methodology ($n = 1$). The final sample comprised 433 datasets extracted from 90 papers.

Reanalysis Procedure

To ensure the results of the adjusted analyses corresponded to the published results, each dataset was first reanalysed by following the procedure adopted by the original authors. All the analyses were performed in R 4.0.5 (R Core Team, 2019). Package *metafor* (Viechtbauer, 2010) was used for fitting the original meta-analytic models. Where the authors did not provide enough information about the fitted model, default settings from *metafor* were used (specifically, random-effects model with REML heterogeneity estimator).

As the majority of the original analyses synthesised the effect sizes into correlation coefficients, r was chosen as the target metric for this study. Studies that reported their estimates in metrics other than the correlation coefficient were first reanalysed using the original metric to ensure the original procedure was being accurately reproduced, and

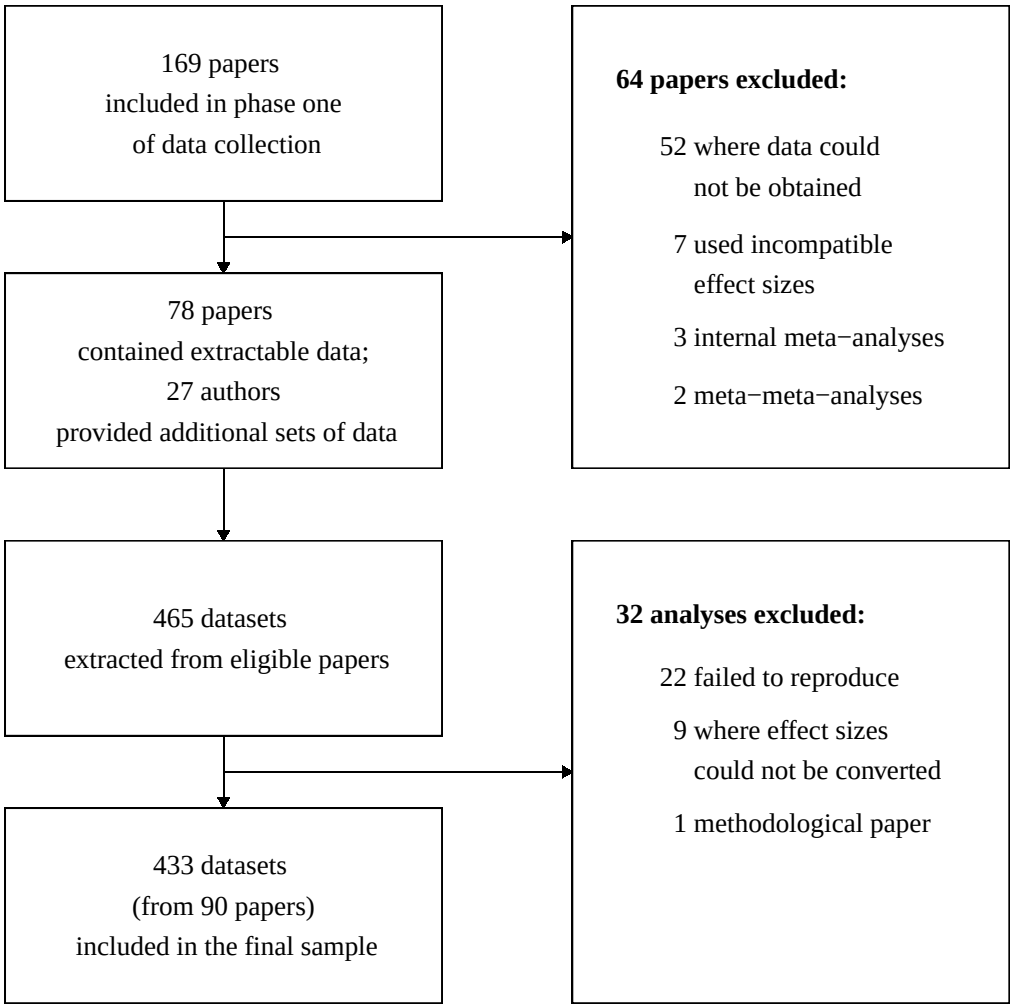

Figure 2

Summary of the sampling process and the reasons for exclusion at different stages of the data extraction and reanalysis.

subsequently converted into r using formulae specified in Borenstein et al. (2011). As both Cohen's d and r can be expressed as the proportion of variance explained by a model (Cohen, 2013), d can be directly converted into r and vice versa with the formula:

$$r = \frac{d}{\sqrt{d^2 + a}} \tag{1}$$

where a is the adjustment for unequal sample sizes defined as:

$$a = \frac{(n_1 + n_2)^2}{n_1 n_2} \tag{2}$$

Effect size metrics for categorical outcomes were first converted into log odds ratios, which were subsequently converted into Cohen's d. This conversion is possible due to the assumption that an underlying continuous trait with logistic distribution exists in each group. The formula used for this conversion was:

$$d = LogOddsRatio \times \frac{\sqrt{3}}{\pi} \tag{3}$$

Further details on conversion among effect sizes can be found in Borenstein et al. (2011). Cohen (2013) discusses the relationship between r and d, while the simulation by Sanchez-Meca et al. (2003) demonstrates the conversion between d and the odds ratio using various formulae, including the one applied here.

For each meta-analysis, we computed the variance explained by the model by squaring the r values and obtaining the R^2. For an analysis to be considered as accurately reproduced, the change in R^2 between the original model and the reanalysed model had to be below 0.03. This value was chosen because (1) a change of 0.03 would not make any practical difference to the conclusions of the analysis and (2) 0.03 was the most conservative value that would also allow for the discrepancies between the original and the reanalysed estimates caused by rounding or conversion of various effect size metrics (like Cohen's d, odds ratios, or

raw differences) into r. Analyses that failed to reproduce the estimates within this limit were not included in the final sample. Next, each of the datasets was assigned appropriate publication bias adjustment methods. The methods most suitable for an individual dataset were assigned based on the findings from Carter et al's (2019) simulation study by evaluating the method performance under conditions typical in meta-analytic research in psychology. These conditions included the number of primary studies in the dataset, level of heterogeneity in the sample, severity of publication bias, and the population effect size.

Level of heterogeneity within datasets was estimated using τ values for random-effects models and Q values for fixed-effects models. Severity of publication bias and the true effect size δ are variables unknown at the sample level. Therefore each dataset was assigned the estimator that was evaluated as best performing given the dataset's sample size and heterogeneity, while assuming (1) moderate publication bias and δ of 0.2, (2) high publication bias and δ of 0.2, (3) moderate publication bias and δ of 0.5, and (4) high publication bias and δ of 0.5. The assumed levels of publication bias (moderate and high) and the population effect size ($\delta = 0.2$ and $\delta = 0.5$) were selected as these values are the most likely to be representative of the data and effects commonly found in psychology research (J Cohen, 1992; Field, 2018; Scheel et al., 2020). This resulted in four estimators being assigned to each dataset, representing four scenarios plausible for psychology research. For each of the scenarios, adjustment method with the lowest mean error (ME) and the lowest root mean square error (RMSE) was selected. ME and RMSE were selected as these are likely to be of interest in situations where meta-analysis is applied. Specifically, ME affects the accuracy of the point estimates and represents the average difference between the simulated estimates and the population estimate. An estimator with an ME that is too high or too low is considered biased. Carter et al. (2019) provide guidance of deciding on an acceptable level of mean error in an estimator for a variety of situations. RMSE contains information about the variability of estimates produced in a simulation. An estimator with an RMSE close to zero will be efficient - it will produce estimates with the least amount of

variance. The aim of a meta-analysis is to produce an accurate parameter estimate with, ideally, the smallest amount of variance. For this purpose, considering the ME and RMSE of an estimator is the most pertinent. Other metrics used for evaluating the performance of estimators include the Type I. error rate and the coverage probabilities of the confidence intervals. If a meta-analyst wishes to perform null hypothesis significance testing, these metrics might be of interest, as well as ME and RMSE. In general however, testing whether an effect is statistically significantly different from zero (or other value that the researcher defines as the null hypothesis), is not the norm in meta-analytic literature. If a method with lowest the ME and RMSE required p values in order to be run but the dataset did not contain them, a method with the second lowest ME and RMSE was selected. Further technical details of method assignment are outlined in the preregistration summary (https://osf.io/kxjs3). Once the appropriate methods were assigned, four alternative models adjusting for the presence of publication bias were fitted to each dataset.

Estimation

Bayesian estimation was used in the fitted models because it is more suitable for addressing the objectives of this study. In the frequentist framework, the point estimates provided by the model are accompanied by p values and confidence intervals constructed around the estimate. The p value is the probability of observing a statistically significant effect equal to or greater than the observed effect assuming the null hypothesis is true (in this context, the hypothesis that there is no difference between original and the adjusted estimates). It cannot however provide evidence for the plausibility of an effect of zero, which is a valid possibility in the current study. Additionally, the p value in a large sample can be statistically significant even if the effect under study (i.e. the difference between the original and the adjusted estimates) is negligibly small, and as such can be a misleading metric for quantifying evidence. The 95% confidence interval constructed around a point estimate represents the limits within which the population value will lie in 95% of the samples.

However in 5% of the samples, this interval will not contain the true population value, and there is no way to know whether the obtained confidence interval is one of the 95% that do (Morey et al., 2016). Furthermore, because different adjustment methods were assigned to different datasets with varying frequency (e.g, PET-PEESE could be evaluated as the best performing for more studies than trim-and-fill in a particular scenario), this could create an inaccurate impression that some methods show more variability than others where the width of the interval is greater with decreasing sample size.

For these reasons, adopting a Bayesian frameword offers advantages in situations where zero is a plausible effect (e.g. Field et al., 2020) as it allows for the construction of intervals that can describe the range of plausible values more accurately than the Frequentist confidence intervals. Much like Frequentist models, Bayesian models produce parameter estimates that, in our case, can quantify the difference between the original estimates and the estimates produced by the adjustment methods. The estimates are derived from two components: the data and the prior probability distribution which represents the beliefs about the values of the parameter (McElrath, 2020). The prior distribution is updated using the sample data, producing a posterior probability distribution. The Bayesian estimates are therefore expressed probabilistically - the parameter estimate is the value with the highest posterior probability. Provided the absence of bias in either model, the parameter estimates of the Bayesian model will converge with the estimates from a Frequentist model if the prior distribution is completely uninformative (i.e. the distribution allows a wide range of values for the parameter). Highest Posterior Density intervals (HPD) can be constructed around the point estimates. 95% HPD interval represents the values that the parameter can plausibly take with 95% probability. Unlike Frequentist intervals, HPD intervals make no apriori assumption about the null hypothesis and can help determine whether zero is a plausible value for the difference between the original and the adjusted estimates. As such, Bayesian parameter estimates and the HPD intervals can directly address the objectives of this study.

Data Analysis

To estimate the typical attenuation in published meta-analytic estimates after the application of the alternative methods while accounting for the hierarchical data structure, four robust 3-level Bayesian models were fitted using Markov Chain Monte Carlo estimation method. The models were fitted separately for each of the four scenarios outlined above, resulting in four models in total. Published meta-analytic papers typically report multiple meta-analyses estimating the effects for multiple outcomes. As such the estimates from these datasets are not independent. In addition, multiple estimates were generated for each meta-analysis, representing the values produced by the original and the publication bias adjustment methods. Therefore, the effect sizes (level 1) were nested within meta-analyses (level 2) which themselves were nested within published papers (level 3).

The model took the following form:

Level 1 (effect size)

$$\text{effect size}_{ijk} = \widehat{\beta}_{0jk} + \widehat{\beta}_{1jk}\text{adjustment method}_{ijk} + R_{ijk} \tag{4}$$

Level 2 (meta-analysis)

$$\widehat{\beta}_{0jk} = \widehat{\gamma}_{00k} + \widehat{U}_{0jk}$$
$$\widehat{\beta}_{1jk} = \widehat{\gamma}_{10k} + \widehat{U}_{1jk}$$

Level 3 (paper)

$$\widehat{\gamma}_{00k} = \widehat{\delta}_{000} + \widehat{V}_{0jk}$$
$$\widehat{\gamma}_{10k} = \widehat{\delta}_{100} + \widehat{V}_{1jk}$$

where the effect size is predicted from the type of the adjustment method. For each of the four models, the estimation method variable has eight potential levels - 3PSM, 4PSM, PET,

PEESE, PET-PEESE, WAAP-WLS, trim-and-fill, and no adjustment[3] - dummy coded in a way where each adjustment method is compared against the baseline unadjusted estimate. However, because each meta-analysis was assigned only one adjustment method per model, only methods that were evaluated as best performing were represented in the factor levels of the estimation method variable. In other words, if an adjustment method was not evaluated as suitable for any of the datasets, it was not represented in the model.

Random intercepts ($\widehat{\beta}_{0jk}$, $\widehat{\gamma}_{00k}$) and random slopes ($\widehat{\beta}_{1jk}$, $\widehat{\gamma}_{10k}$) were modelled at level 2 and level 3 which accounted for variability in the magnitude and direction of the effects across the meta-analyses and the papers (Appendix A contains a visual summary of these effects). 95% HPD intervals were extracted from the four models to estimate the plausible range of these effects across the different scenarios and the adjustment methods. The models were fitted with the *brms* package (Bürkner, 2017). Default priors from this package were used as there were no prior expectations that would justify the use of more informative prior distributions. This included improper flat priors (unconstrained uniform priors) set for $\widehat{\beta}_{1jk}\text{adjustment method}_{ijk}$, and a Student's t priors set for the intercept $\widehat{\beta}_{0jk}$ and for the σ parameter (the residual standard deviation). As such the estimation was entirely data driven. Link functions from the Student family were specified for the response distribution. The use of the Student's t response distribution (as opposed to the default Gaussian distribution) allows estimation that is robust to violations of the assumptions of general linear models, like the presence of heteroscedasticity and outliers (Gelman & Hill, 2006) which are commonly found in statistical models within psychology (Field & Wilcox, 2017; Wilcox, 2016). To check the robustness of the findings we conducted a sensitivity analysis comparing all the fitted models with Bayesian models where the priors for all parameters in the model were set

[3] The original intention was to also assess p-curve and p-uniform, however after further assessment of the sample, p-uniform was evaluated as well performing for only four datasets (p-curve for zero), none of which contained the parameters necessary to compute an estimate adjusted by this method. p-curve and p-uniform were therefore dropped from the present study.

to a uniform distribution constrain between -2 and 2. We also compared the results from each model with a corresponding Frequentist model fitted using modified robust *M*-estimation (Koller & Stahel, 2011). This estimation was chosen because unlike the commonly used maximum likelihood (ML) or the restricted maximum likelihood (REML) estimators, *M*-estimators are robust to violations of statistical assumptions and the presence of extreme cases, and as such provide a more suitable comparison for models using the MCMC estimation.

Preregistration of this study can be found at https://osf.io/kxjs3, and a document outlining where the analysis diverged from the preregistered plan is available at https://osf.io/k9hqm.[4] This study was approved by the University of Sussex Psychology School Research Ethics Officers.

Results

Data Screening

After the application of the adjustment methods, some effect size estimates exceeded the boundaries of the correlation coefficient. Any estimates for which the 95% confidence intervals went below -1 or above 1 were excluded from the analysis, resulting in the exclusion of 43 cases from Model 1, 46 cases from Model 2, 19 cases from Model 3, and 29 cases from Model 4. These cases are summarised at the end of the section. Additionally, some adjustment methods were not represented in a sufficient number of cells, and therefore could not be included in the models as predictor levels. For Model 1, 4PSM was only applied in 5 cases. Because 3PSM is considered a fallback option of 4PSM (Carter et al., 2019), these two

[4] In addition, an alternative analysis was suggested by a reviewer where all the adjustment methods are fitted to all of the datasets. We report this analysis in a supplemental document, and the relevant data and R code are available at https://osf.io/k9hqm

methods were merged into a single category for this Model. For Model 2, cases that used WAAP-WLS ($n = 2$) were excluded. For Model 3, cases that used PET ($n = 4$) were excluded. For all the exclusions of the adjusted estimates, their unadjusted counterparts were also excluded from the final conditions

Sample Description

The final sample consisted of 774 cases for Model 1, 764 for Model 2, 816 for Model 3, and 806 for Model 4. Median number of primary studies included in analyses was 14, ranging from 2 to 752. The median τ value was 0.11, ranging from 0 to 0.60. Therefore, the present sample was comparable to the range of k and τ values simulated by Carter et al. (2019). Table 2 contains the sample summary values of k and τ for the different adjustment methods across the four models.

Sampling Diagnostics

Samples were drawn using the No-U-Turn Sampler (Hoffman & Gelman, 2014). To assess the performance of the sampler, we looked at the potential scale reduction statistics $\hat{R}$, post-warm-up divergent transitions, and bulk and tail effective sample sizes. Sampling diagnostics were performed using the *shinystan* package for diagnostic calculations (Gabry, 2018). For all four models, the $\hat{R}$ values were below 1.01, indicating that chains within each model converged to common distribution (Vehtari et al., 2020). No post-warm-up iterations for predictors across all four models encountered divergent transitions. Bulk and Tail Effective Sample Sizes for all predictors exceeded the minimum recommended threshold of 100 per Markov Chain (400 per model) indicating reliability of posterior point estimates and their HPD intervals (Vehtari et al., 2020). Full diagnostic report, including visual diagnostics, can be accessed at https://osf.io/k9hqm/.

Table 2

Median k and τ values for the adjustment methods across the four conditions

	Moderate PB, $\delta = 0.2$		High PB, $\delta = 0.2$		Moderate PB, $\delta = 0.5$		High PB, $\delta = 0.5$	
Adjustment method	k	τ	k	τ	k	τ	k	τ
3PSM	28.0	0.127	28.0	0.134	28.0	0.124	24.5	0.130
4PSM	28.0	0.127	26.0	0.155	27.0	0.162	46.0	0.137
PET	26.0	0.105	24.0	0.094			25.0	0.140
PEESE	7.0	0.042	8.0	0.015	10.0	0.118	15.5	0.087
PET-PEESE	9.5	0.147	9.5	0.147	24.0	0.140	9.5	0.147
WAAP-WLS	33.0	0.043			7.0	0.015	6.0	0.002
TF							32.0	0.039

Note. PB: publication bias; 3PSM: three-parameter selection model; 4PSM: four-parameter selection model; PET: precision-effect test; PEESE: precision-effect estimate with standard error; WAAP-WLS: weighted average of adequately powered studies - weighted least squares; TF: trim-and-fill. 3PSM and 4PSM were merged into a single category for the first condition (moderate publication bias, $\delta = 0.2$). Where values are missing, the method was not evaluated as best performing for any of the datasets in a given condition

Model 1: Moderate Publication Bias, $\delta = 0.2$

Random Effects

At both level 2 (analysis) and level 3 (paper level) the modelled relationships between the effect size and the type of estimator showed variance in the intercepts, $SD_{L_2} = 0.099$, 95% HPD [0.089, 0.11], $SD_{L_3} = 0.106$ [0.083, 0.129], and slopes, $SD_{L_2} = 0.015$ [0.003, 0.029], $SD_{L_3} = 0.027$ [0.014, 0.04], suggesting that the magnitude and the direction of the relationship varied across analyses and papers. The extent of this variance was small as evident from the point estimates and HPD intervals for slopes, which provide the range of plausible population values for this estimate with 95% probability. At both levels, the correlation between intercepts and slopes ranged from small to large, $corr_{L_2} = 0.593$ [0.157, 1], $corr_{L_3} = 0.243$ [-0.156, 0.612].

Table 3

Average effect sizes (r) for publication bias adjustment methods for Scenario 1

Adjustment method	Effect size (r)	Standard error	2.5% HPD	97.5% HPD
None	0.215	0.013	0.189	0.242
3PSM + 4PSM	0.221	0.015	0.209	0.233
PET	0.200	0.017	0.180	0.220
PEESE	0.227	0.016	0.212	0.241
PET-PEESE	0.267	0.018	0.246	0.289
WAAP-WLS	0.211	0.019	0.183	0.237

Note. Estimated effect sizes account for the hierarchical data structure. Each effect size is adjusted for the presence of the other dummy-coded predictor levels in the model. 3PSM and 4PSM were merged into a single level. Trim-and-fill was not evaluated as performing well under the conditions of Scenario 1 and therefore is not represented in Model 1.

Fixed Effects

Table 3 and Figure 3 show average effect sizes for each estimator evaluated in Model 1. Mean effect size for the baseline unadjusted estimate (intercept) was 0.215 [0.189, 0.242]. Effect size attenuation occurred for PET, b = -0.015 [-0.035, 0.005], and WAAP-WLS, b = -0.004 [-0.032, 0.022]. The HPD intervals included zero as a plausible population value. All other methods tended to adjust the point estimate upwards, with the largest upward adjustment shown by PET-PEESE, b = 0.052 [0.03, 0.074], suggesting that on average, correlation coefficients estimated by PET-PEESE tended to be greater than their unadjusted counterparts by 0.052 units. Table 4 contains the summary of all fixed effects *beta* estimates and their HPD intervals across the four models.

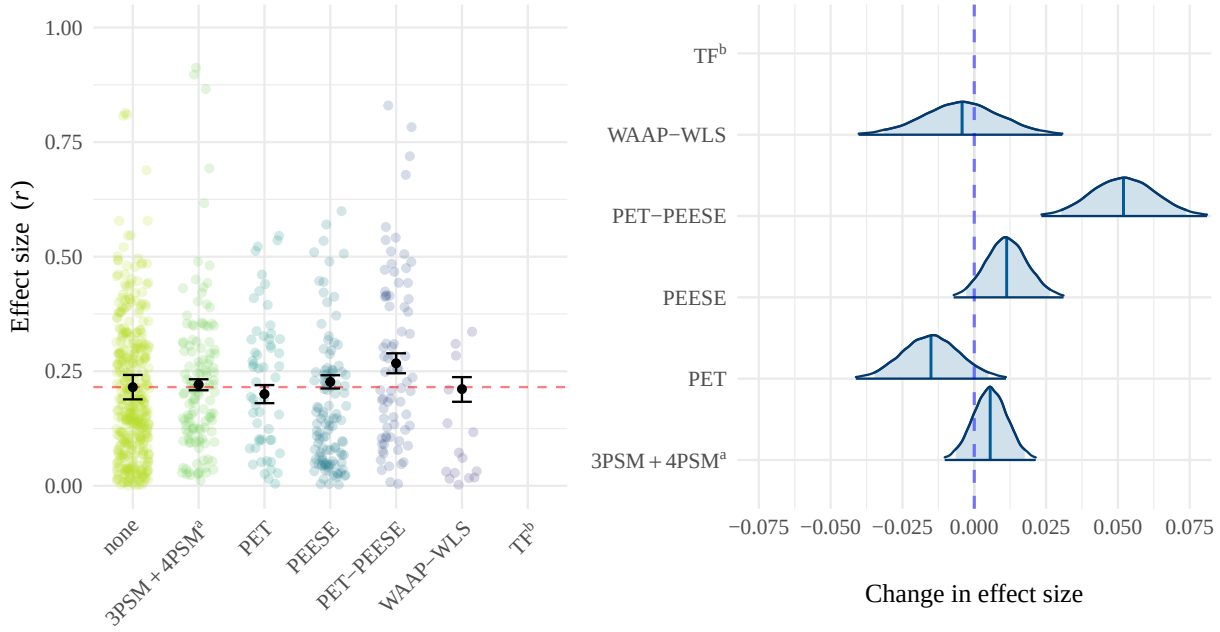

Figure 3

Adjusted means and 95% HPD intervals for publication bias adjustment methods in Model 1 (left), and posterior density plots of change in effect size when compared against the unadjusted baseline (right). Note. *a. 3PSM and 4PSM were merged into a single factor level. b. Trim-and-fill was not evaluated in Model 1.*

Table 4

Summaries of beta estimates for all four models.

	Model 1		Model 2		Model 3		Model 4	
term	*b*	95% HPD	*b*	95% HPD	*b*	95% HPD	*b*	95% HPD
intercept	0.215	[0.189, 0.242]	0.216	[0.190, 0.243]	0.214	[0.188, 0.241]	0.215	[0.189, 0.243]
3PSM	0.006	[–0.007, 0.017]	0.007	[–0.009, 0.023]	0.006	[–0.004, 0.015]	0.009	[–0.003, 0.021]
4PSM	0.006	[–0.007, 0.017]	0.017	[–0.017, 0.052]	0.034	[0.006, 0.061]	–0.006	[–0.030, 0.017]
PET	–0.015	[–0.035, 0.005]	–0.001	[–0.018, 0.016]			–0.032	[–0.055, –0.009]
PEESE	0.011	[–0.003, 0.026]	0.003	[–0.013, 0.019]	0.021	[0.010, 0.032]	0.005	[–0.007, 0.017]
PET-PEESE	0.052	[0.030, 0.074]	0.055	[0.033, 0.078]	–0.022	[–0.043, –0.001]	0.052	[0.034, 0.072]
WAAP-WLS	–0.004	[–0.032, 0.022]			0.002	[–0.007, 0.012]	0.000	[–0.010, 0.009]
TF							0.006	[–0.011, 0.022]

Note. Betas represent change in the effect size (r) after the application of an alternative method. 3PSM and 4PSM were merged into a single category for Model 1. Where values are missing, the method was not evaluated as best performing for any of the datasets in a given model.

Model 2: High Publication Bias, $\delta = 0.2$

Random Effects

The intercepts varied at level 2 and level 3, $SD_{L_2} = 0.098$ [0.087, 0.11], $SD_{L_3} = 0.105$ [0.083, 0.129], and there was also small variation in magnitudes and directions of the effect across analyses and papers, $SD_{L_2} = 0.021$ [0.007, 0.038], $SD_{L_3} = 0.025$ [0.008, 0.041]. At both levels, the correlation between intercepts and slopes ranged from small to large, $corr_{L_2} = 0.611$ [0.179, 1], $corr_{L_3} = 0.170$ [-0.342, 0.639], however the HPD intervals indicated that zero as well as a small negative correlation are plausible true values for the correlation at level 3.

Table 5

Average effect sizes (r) for publication bias adjustment methods for Scenario 2

Adjustment method	Effect size (r)	Standard error	2.5% HPD	97.5% HPD
None	0.216	0.014	0.190	0.243
3PSM	0.223	0.008	0.207	0.238
4PSM	0.233	0.018	0.199	0.268
PET	0.215	0.009	0.197	0.232
PEESE	0.219	0.008	0.203	0.235
PET-PEESE	0.271	0.012	0.249	0.294

Note. Estimated effect sizes account for the hierarchical data structure. Each effect size is adjusted for the presence of the other dummy-coded predictor levels in the model. Trim-and-fill was not evaluated as performing well under the conditions of Scenario 2 and therefore is not evaluated in Model 2. WAAP-WLS was excluded as it was not represented in sufficient number of cells.

Fixed Effects

Table 5 and Figure 4 show average effect sizes for each estimator evaluated in Model 2. Mean effect size for the baseline unadjusted estimate (intercept) was 0.216 [0.19, 0.243]. Effect size attenuation occurred only PET , b = -0.001 [-0.018, 0.016], however the magnitude of this effect size was small and the HPD intervals indicated that zero was within the plausible range of values for this effect. All other methods tended to adjust the estimate upwards, with the largest upward adjustment shown by PET-PEESE, b = 0.055 [0.033, 0.078], suggesting that on average, correlation coefficients estimated by PET-PEESE tended to be greater than their unadjusted counterparts by 0.055 units. Table 4 contains the fixed-effects summaries.

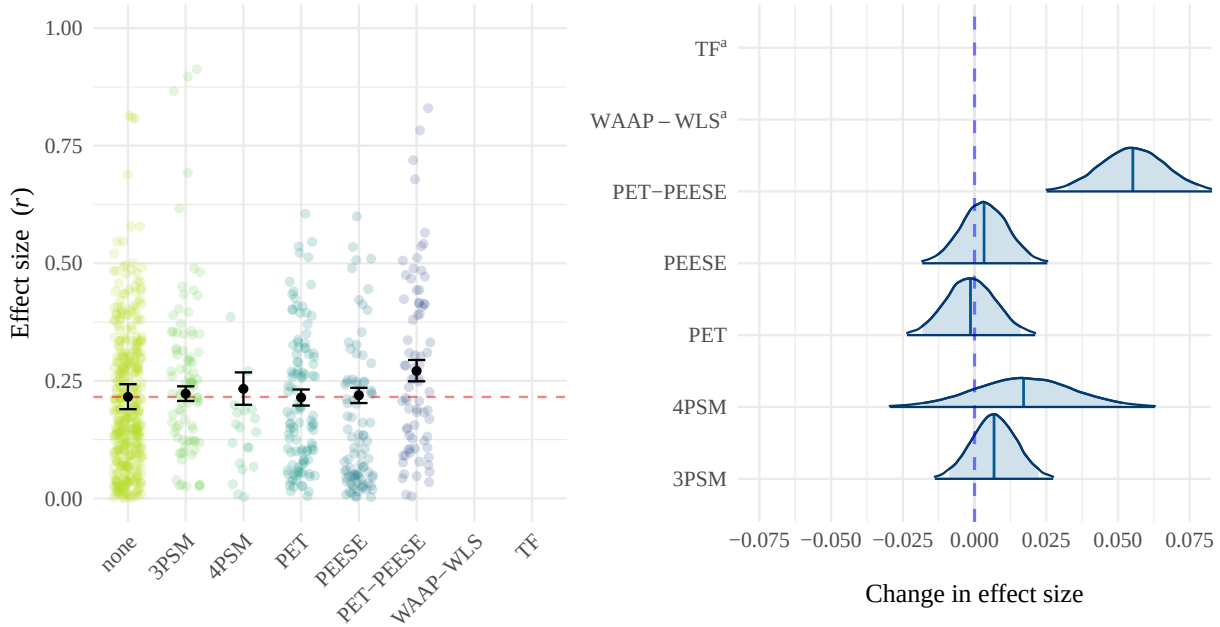

Figure 4

Adjusted means and 95% HPD intervals for publication bias adjustment methods in Model 2 (left), and posterior density plots of change in effect size when compared against the unadjusted baseline (right). Note. *a. Trim-and-fill and WAAP-WLS were not evaluated in this model.*

Model 3: Moderate Publication Bias, $\delta = 0.5$

Random Effects

The intercepts varied at level 2 and level 3, $SD_{L_2} = 0.101$ [0.092, 0.11], $SD_{L_3} = 0.107$ [0.086, 0.129]. There was a minimal variation in magnitudes and directions of the effects across papers, $SD_{L_3} = 0.027$ [0.018, 0.036], and on the analysis level, $SD_{L_2} = 0.004$ [0, 0.008]. At both levels, the correlation between intercepts and slopes ranged from small to large, $corr_{L_2} = 0.610$ [-0.044, 1], $corr_{L_3} = 0.325$ [0.014, 0.625], however the HPD intervals indicated that zero as well as a small negative correlation are plausible true values for the effect at level 3.

Table 6

Average effect sizes (r) for publication bias adjustment methods for Scenario 3

Adjustment method	Effect size (r)	Standard error	2.5% HPD	97.5% HPD
None	0.214	0.013	0.188	0.241
3PSM	0.220	0.005	0.210	0.229
4PSM	0.249	0.014	0.220	0.275
PEESE	0.235	0.006	0.224	0.246
PET-PEESE	0.192	0.011	0.171	0.213
WAAP-WLS	0.216	0.005	0.207	0.225

Note. Estimated effect sizes account for the hierarchical data structure. Each effect size is adjusted for the presence of the other dummy-coded predictor levels in the model. Trim-and-fill was not evaluated as performing well under the conditions of Scenario 3 and therefore is not evaluated in Model 3. PET was excluded as it was not represented in sufficient number of cells.

Fixed Effects

Table 6 and Figure 5 show average effect sizes for each estimator evaluated in Model 3. Mean effect size for the baseline unadjusted estimate (intercept) was 0.214 [0.188, 0.241]. Effect size attenuation occurred only for PET-PEESE, b = -0.022 [-0.043, -0.001]. All other methods tended to adjust the estimate upwards, with the largest upward adjustment shown by 4PSM, b = 0.034 [0.006, 0.061], suggesting that on average, correlation coefficients estimated by 4PSM tended to be greater than their unadjusted counterparts by 0.034 units. Table 4 contains the fixed-effects summaries.

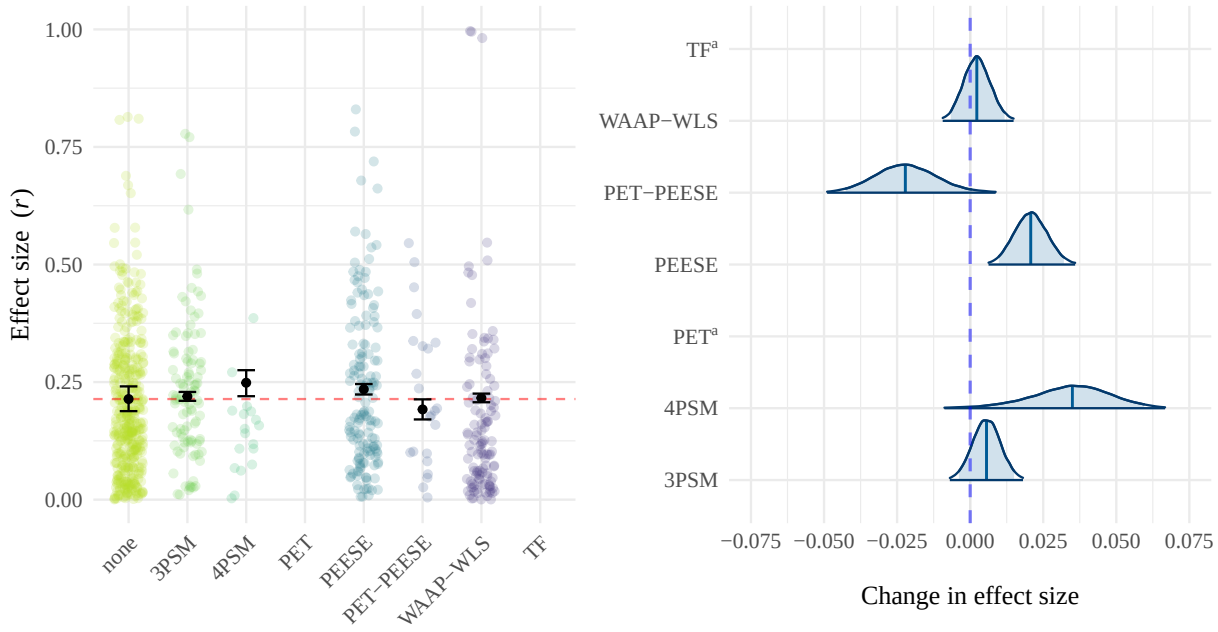

Figure 5

Adjusted means and 95% HPD intervals for publication bias adjustment methods in Model 3 (left), and posterior density plots of change in effect size when compared against the unadjusted baseline (right). Note. *a. PET and Trim-and-fill were not evaluated in this model.*

Model 4: High Publication Bias, $\delta = 0.5$

Random Effects

The intercepts varied at level 2 and level 3, $SD_{L_2} = 0.100$ [0.091, 0.109], $SD_{L_3} = 0.109$ [0.087, 0.131]. There was a minimal variation in magnitudes and directions of the effect across papers, $SD_{L_3} = 0.022$ [0.011, 0.033], and at the analysis level, $SD_{L_2} = 0.003$ [0, 0.008]. At level 2, the plausible values for the correlation between intercepts and slopes ranged from a large negative to a large positive effect, $corr_{L_2} = 0.351$ [-0.598, 1], and at level 3, the plausible values for this correlation ranged from small negative to large positive correlation, $corr_{L_3} = 0.340$ [-0.05, 0.722], including zero as a plausible population value.

Table 7

Average effect sizes (r) for publication bias adjustment methods for Scenario 4

Adjustment method	Effect size (r)	Standard error	2.5% HPD	97.5% HPD
None	0.215	0.014	0.189	0.243
3PSM	0.224	0.006	0.212	0.236
4PSM	0.210	0.012	0.185	0.232
PET	0.183	0.012	0.160	0.206
PEESE	0.220	0.006	0.208	0.232
PET-PEESE	0.267	0.010	0.249	0.287
WAAP-WLS	0.215	0.005	0.205	0.224
TF	0.221	0.008	0.204	0.237

Note. Estimated effect sizes account for the hierarchical data structure. Each effect size is adjusted for the presence of the other dummy-coded predictor levels in the model.

Fixed Effects

Table 7 and Figure 6 show average effect sizes for each estimator evaluated in Model 4. Mean effect size for the baseline unadjusted estimate (intercept) was 0.215 [0.189, 0.243]. Effect size attenuation occurred for PET, b = -0.032 [-0.055, -0.009], and 4PSM, b = -0.006 [-0.03, 0.017], the latter indicating zero as a plausible value. Point estimate for WAAP-WLS showed little to no difference compared to the baseline unadjusted estimate $b = 0$ [-0.01, 0.009]. All other methods tended to adjust the estimate upwards, with the largest average upward adjustment shown by PET-PEESE, $b = 0.052$ [0.034, 0.072]. Table 4 contains the fixed-effects summaries.

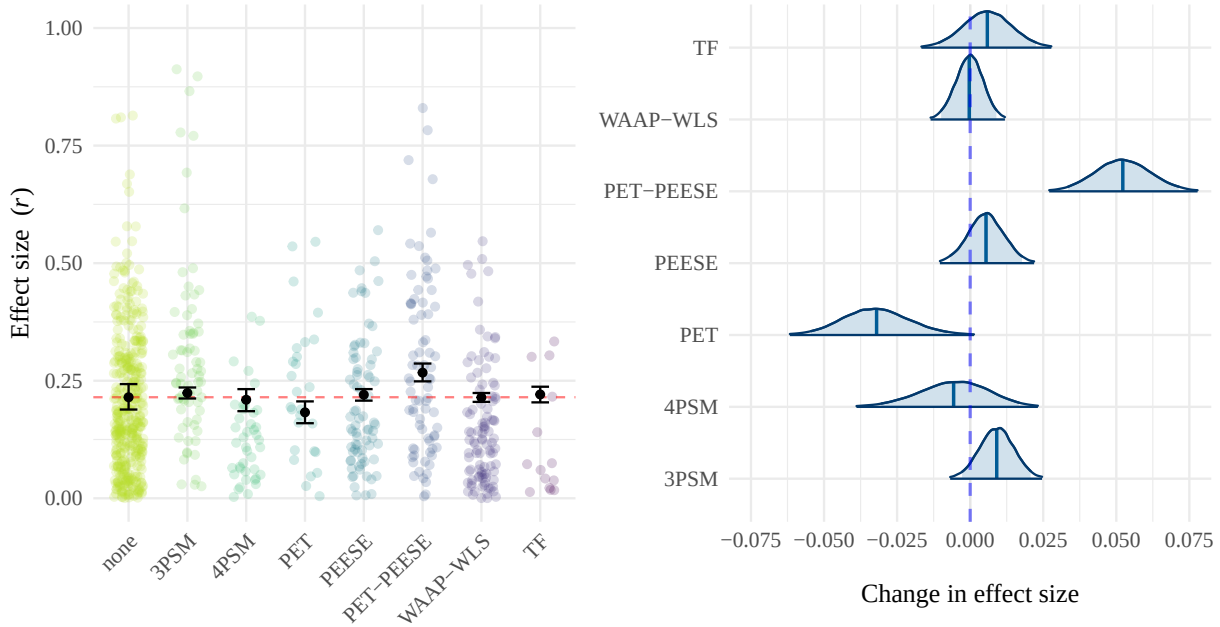

Figure 6

Adjusted means and 95% HPD intervals for publication bias adjustment methods in Model 4 (left), and posterior density plots of change in effect size when compared against the unadjusted baseline (right).

Excluded Cases

A number of cases were excluded (126 analyses from 26 papers, across the four models) as the adjusted estimates or their confidence intervals crossed the minimum and maximum boundaries of the correlation coefficient. These cases are summarised in Table 8. The estimators that were most frequently represented in the excluded sample were PET-PEESE ($n = 66$), PEESE ($n = 43$), and PET ($n = 8$). Median sample size for these methods ranged from 3 to 18, with the exception of 2 PET analyses from a single paper with $k = 232$. Median heterogeneity levels for PET, PEESE and PET-PEESE ranged from moderate to high. Cases using 3PSM ($n = 1$) and 4PSM ($n = 3$) had a median sample size of $k = 27$ and showed high heterogeneity levels ($Mdn\ \tau = 0.429$). Five cases in the excluded sample used WAAP-WLS and their sample sizes ranged from $k = 2$, to and $k = 25$. Heterogeneity levels for all WAAP-WLS samples were close to zero. Almost all the estimators tended to adjust the effect size upwards, with the most extreme change in r observed in 3PSM (median change of +0.557) and 4PSM (+0.361), except for WAAP-WLS, which tended to attenuate the effect size (–0.049).

Table 8

Descriptive summary of the sample excluded on the basis of implausible adjusted values

Adjustment method	n	Mdn sample size(k)	Min k	Max k	$Mdn\ \tau$	Mdn change in r
3PSM	1	27	27	27	0.429	0.557
4PSM	3	27	27	27	0.429	0.361
PET	8	120	3	233	0.081	0.332
PEESE	43	3	3	6	0.025	0.151
PET-PEESE	66	4	3	18	0.196	0.32
WAAP-WLS	5	16	2	25	0.000	–0.049

Sensitivity Analysis

The results of the models were consistent across the models using the different estimation methods. The sensitivity analysis revealed no notable differences between the models using the default improper flat priors and the models using uniform priors constrained between -2 and 2. The point estimates and the HPD intervals differed only on the fourth decimal place. The estimates from the Bayesian models were also consistent with the estimates from the robust Frequentist models. The differences between the Frequentist *M*-estimates and the sets of Bayesian estimates were only on the third decimal place. As such, a selection of an alternative estimation approach would not have altered the conclusions of the statistical analysis presented here. The confidence intervals aligned with the HPD intervals in majority of the cases. Discrepancies on the second or third decimal place were identified in cases where an adjustment method was assigned as the most appropriate with lower frequency compared to the other methods (specifically, 4PSM and trim-and-fill), which resulted in slightly wider intervals for these methods. As explained in the Estimation section, this difference is to be expected because of the fundamental differences in the Bayesian and Frequentist estimation approaches. As the differences across the models in the sensitivity analysis were minimal, we do not report these in the main text and instead provide the results as part of the supplemental materials (https://osf.io/k9hqm).

Discussion

Publication bias can undermine the attempts to statistically synthesise effects across studies with accuracy. This study evaluated how applying appropriate publication bias adjustment methods changes the effect size estimates of published meta-analyses. The typical adjustment - and the range of this adjustment - found for the different methods was small, making a difference on the second or third decimal place when compared to the unadjusted estimate. The direction of the adjustment of point estimates varied, where some

methods tended to adjust the effect size slightly upwards. However, where this happened, the lower bound of the 95% HPD interval always indicated downward adjustment as also plausible. An exception to this trend was PET-PEESE. In three out of four models, the upward adjustment of PET-PEESE was greater when compared to the other methods. This effect can be understood by examining the excluded studies. 66 cases that used PET-PEESE as the adjustment method had to be excluded from the statistical models because the adjusted estimates or their confidence intervals crossed the limits of the correlation coefficient. This set of excluded cases showed a large average upward adjustment (+.32), the common denominators being small sample sizes (with less than ten primary studies) with large heterogeneity. Further inspection of the data showed that cases with only slightly larger sample sizes made it through the filter and into the analysis, but still showed a large upward adjustment, dragging the overall model estimates upwards. The data showing this effect can be accessed at https://osf.io/k9hqm/. The characteristics of the excluded cases (small sample sizes or high levels of heterogeneity) indicate that even if a method is evaluated as performing well on average in simulation studies under specific sample-level and population-level characteristics, it may still underperform when applied to real data. Upward adjustments should be viewed as an indicator of unsuitability of the selected method for the dataset at hand, rather than as evidence of no publication bias.

When the attenuation in the estimates did occur, the magnitude of the adjustment was mostly consistent across the four models and in line with the extent that would be expected based on the simulation studies when scaled down to correlation coefficients (as opposed to typically simulated uncapped effect sizes like Cohen's d). Where the attenuation was observed, the magnitude of the adjustment would typically not be enough to warrant an alteration in the way the meta-analysed effect is interpreted. This finding is in line with a study by Van Aert et al. (2019), who found that after applying publication bias adjustment methods to homogeneous datasets which mitigated the bias in performance associated with heterogeneity, the change in the effect sizes from the unadjusted estimates was minimal. Van

Aert et al. (2019) discuss this finding as an indicator of low levels of publication bias in the field. This conclusion is in stark contrast with studies that investigate publication bias by the means of looking at the proportion of published statistically significant results. Studies consistently report that such results are over-represented in the published literature with over 90% of papers reporting positive results (Fanelli, 2010; Scheel et al., 2020), which is at odds with the typically observed power in psychological research of 50% (Cohen, 2013; J. Cohen, 1992).

While it is true that estimates adjusted for publication bias are commonly used for sensitivity analysis (Carter et al., 2019; Kim et al., 2014; Vevea & Woods, 2005), any identified difference between the adjusted and the original estimate does not contain information which would be indicative of the extent of the bias in the field. All the methods discussed thus far adjust the estimate based on what is present in the sample. Therefore, the methods are using a likely biased sample to compensate for what is missing. This might be a valid approach to obtain an estimate more representative of the population effect, however it is not an accurate way to estimate the extent of publication bias in the field. Figure 7 illustrates this point. The figure displays a randomly selected batch of estimates from Carter et al.'s simulation plotted against their sample sizes. The plotted population effect size is $r = .10$ (equivalent to $\delta = 0.2$ to when the group sample sizes are equal)[5] with high levels of heterogeneity. Assuming an extreme version of publication bias where statistically non-significant studies never get published and included in meta-analyses, the difference between the population estimate and the biased estimated derived from studies which reported $p < .05$ is approximately 0.09. Of course, the real model of publication bias is more complicated, where the negative effect sizes might also not get published, while some of the non-significant results either do get published or the meta-analyst manages to get hold of the unpublished data. The key point is that even though the change in the estimate is relatively

[5] As pointed out by a reviewer, in a scenario of unequal sample sizes where $n_1 = 50$ and $n_2 = 200$, δ of 0.2 is equivalent to an r of approximately 0.08.

small, the proportion of non-significant and likely unpublished studies is large. Therefore. inferring the extent of publication bias in the field from the extent of attenuation in estimates is misleading. Likewise, the small changes in estimates should not be taken as an invitation to disregard the problem of publication bias altogether. The large proportion of unpublished effect sizes that never make it into meta-analyses may contain relevant information about moderators which likely exist considering the high levels of heterogeneity typically found in meta-analytic samples in psychology (Van Erp et al., 2017).

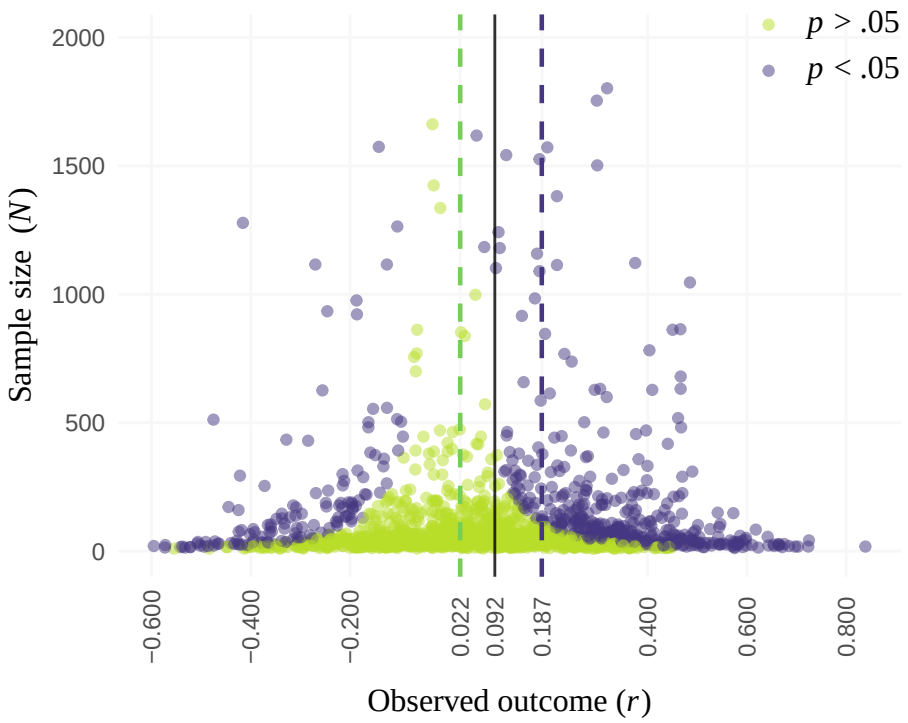

Figure 7

A simulated population of primary studies. Dashed lines represent random-effects meta-analytic estimates for a sample comprised of only statistically significant studies (green or light grey line on the left) and a sample of non-significant studies (purple or dark grey line on the right). Full line represents the population effect size. Plot data were obtained from https://osf.io/rf3ys/ (Carter et al., 2019) .

Although publication bias is an issue pervasive across the whole field and cannot be

fully addressed on the level of an individual meta-analysis, there are steps that researches can take to reduce its impact on analyses, and bring attention to the censoring that is likely present in their sample. We first discuss recommendations based on the results of the present study, and then we offer an overview of some recent developments in the area publication bias adjustment methods that researchers should also consider. This is followed by a wider discussion of how publication bias can be addressed on the field level with reference to the recent methodological reforms in psychology.

Inclusion of Unpublished Studies

As a starting point, researchers should aim to thoroughly search for unpublished literature. The previous phase of the present study showed that 56% of papers did not include any unpublished studies, and when they did, the unpublished studies comprised on average only 6% of the meta-analytic sample (unpublished data - Avery et al., 2020). Importantly, this should not be limited to studies retrieved from their own file drawers, as this can introduce further bias (Ferguson & Brannick, 2012). Admittedly, researchers' own studies can be a representative sample of unpublished studies in particularly narrow research fields, however where possible, it is crucial to extend the literature search beyond conveniently available samples.

Use of Publication Bias Adjustment Methods

Although the attenuation in estimates after applying adjustment methods detected in the present study and in previous research (Van Aert et al., 2019) was minimal, the meta-analysts should still seek to explore the range of plausible effects by applying appropriate adjustment methods while considering different population level-assumptions (like the severity of publication bias, or the population effect size). Discussions around effect sizes in meta-analyses tend to overwhelmingly focus on point estimates with lack of regard for the variance around these estimates. Considering the full range of the plausible

population effects can produce more transparent discussions that are sensitive to the levels of uncertainty around publication bias present in the literature.

Use of the Adjustment Methods with Adequate Sample Sizes

The final recommendation is a cautionary note on the use of publication bias adjustment methods in meta-analyses with extremely small samples ($k < 10$). Small sample sizes were associated with large upward adjustments, even though the methods that were applied were evaluated as well performing given the conditions of the datasets they were applied to. While meta-analysis can be performed with as few as two primary studies (Valentine et al., 2010), researchers should note that for such small samples, the methods outlined in the present study will likely not be able to adequately combat publication bias.

Future Directions in Addressing Publication Bias

Recent Developments in Adjustment Methods

Publication bias adjustment is still an area under active development. A promising set of methods which has seen recent upgrades as well as accessible software implementation include Bayesian Model-Averaged Meta-analysis (BMA; Bartos et al., 2020; Gronau et al., 2017; Guan & Vandekerckhove, 2016; Maier et al., 2020). The main principle of BMA is applying multiple models simultaneously depending on which assumptions each model makes about the data-generating process. For example models could assume the absence or presence of the true population effect, and absence or presence of between study heterogeneity. Therefore, four models in total would be entered into BMA - 2 (null vs alternative hypotheses) $\times 2$ (fixed vs random effects). As a default, all models are assumed to be equally plausible at the beginning. This plausibility is then updated based on the Bayes theorem - models that predict the data well based on the posterior probability are assigned greater plausibility weights than models that predict the data poorly. The models are

subsequently averaged accordingly - the models with greater weights have greater contribution to the overall meta-analytic estimate and vice-versa for models with smaller weights (Gronau et al., 2017). This principle can be extended to include models making different assumptions about publication bias. Maier et al. (2020) incorporated the selection models (McShane et al., 2016; Vevea & Woods, 2005) applied in this study as part of the model set entered into the BMA. Selection models make assumptions about the probability of a publication for each primary study depending on the associated p values, where p values smaller than 0.05 are considered to have higher probability of publication. To account for unpublished studies missing from the meta-analytic sample, the selection models increase the weight of the statistically non-significant effects when computing the meta-analytic estimate. This assumption can also be refined by creating and additional step where "marginally significant" results in the bracket of 0.05 - 0.10 are assigned their own probability. The selection models are entered into BMA along with models assuming no bias and weighted according to how well they predict the data. Maier et al. (2020) and Bartos et al. (2020) offer a detailed description of the method, including applied examples.

This approach offers a number of advantages. Maier et al. (2020) focused on selection models because of their relatively good performance under heterogeneity (Table 1), but BMA is flexible enough to accommodate any of the methods discussed in this paper. Researchers have the option to specify their own prior distributions if they have additional knowledge about the data-generating process, however this is not a requirement. In cases where informative priors are specified, robustness checks with alternative priors should also be included. The methodology also aligns with the recommendations made about frequentist methods for addressing publication bias presented in Carter et al. (2019) and reiterated here - that given the uncertainty around population level characteristics like publication bias or true effect sizes, a single point estimate assuming specific values for these characteristics is likely to be misleading. BMA allows the researchers to explore the plausible range of effect sizes under different assumptions about the data generating process. It also allows a

synthesis into a meta-analytic estimate that incorporates the estimates from different models based on their plausibility.

As with any method, BMA has its weaknesses and brings unique challenges that the applied researchers need to consider. One of the key weaknesses of the methods presented in this paper were small samples with $k < 10$. BMA shows an improvement in the accuracy of estimates in small samples compared to the methods presented here, however its performance can still be suboptimal in this situation (Maier et al., 2020). Real effects can be underestimated if the model assuming the null effect is included, even it has low posterior probability and is down-weighted as a result. Maier et al. (2020) and Bartos et al. (2020) recommend not including the null hypothesis model if the aim is estimation and not hypothesis testing, which alleviates the problem. This issue however highlights another challenge which is the selection of the appropriate models to include in the BMA. Inclusion of a model with poor posterior probability can contaminate the overall estimate, whereas exclusion of a relevant model means that a potentially crucial model is not weighing in on the final estimate. At the moment, more guidance is needed about the nuances of application that would maximise the benefits of BMA.

Wider Efforts to Address Publication Bias

Based on the results and the discussion points presented here, it is clear that there is no single adjustment method that would adequately mitigate publication bias in all the possible situations that the researchers are likely to encounter in applied settings. We discussed a number of strategies than can be applied to make the best use of these methods, providing a direct action that individual researchers can take to address the situation at hand, at least to an extent. However the very existence of publication bias points to wider problems in the field of psychology that cannot be solved by applying *post hoc* solutions. Recent efforts aimed at improving the reproducibility and credibility of scientific findings have brought about a number of methodological reforms that could directly contribute to

creating a less biased landscape of findings in psychology (Chambers, 2019; Open Science Collaboration, 2015). One such reform was the introduction of preregistration, where the researchers their analysis plan prior to collecting or inspecting the data and create a time-stamped, uneditable registration document (Nosek et al., 2019). Preregistration does not solve publication bias but publicly archived analyses plans will help the meta-analysts locate information about studies that may not have made it through the publication process. Registered Reports (Nosek & Lakens, 2014) directly combat publication bias arising from the nature of the produced findings, whether this is generated by editors or reviewers rejecting manuscripts with statistically non-significant results, or researchers deciding not to submit their non-significant results in the first place. For a registered report publication, the researchers first submit their hypotheses, methods, and analysis plans which are reviewed and, if necessary, refined based on feedback. If the peer-review is successful, the journal commits to publishing the paper in principle regardless of the results, as long as the agreed method and analysis plan are carried out.

There is some evidence showing that registered reports have been, so far, successful in producing a relatively balanced scientific record that is more likely to accurately reflect the reality. Scheel et al. (2020) found that the proportion of hypothesis-confirming findings in registered reports is 43.66% which is in a stark contrast with 96.05% hypothesis-confirming results reported in articles taking the standard publication route. Similar findings are reported in Allen and Mehler (2019). At the time of writing, 288 journals provide the opportunity to submit a registered report (https://www.cos.io/), and this number is increasing. In addition, the Open Science Framework (OSF) enables the researchers to preregister analysis plans and store their data and research materials in a an online repository that is part of a searchable database. The practices proposed by the methodological reform play an important role in the efforts to improve the credibility of psychology. It is therefore crucial to build incentive structures that support senior researchers to switch their practice, and encourage the early career researchers to adopt these practices from the get go.

Limitations and Additional Considerations

The present study is limited in a number of respects. Even though each adjustment method was selected for specific datasets because it was evaluated as best performing, some methods still showed varying levels of baseline bias in their mean error (ME) and root mean square error (RMSE). The baseline ME of each adjustment method was however always lower than ME the unadjusted random-effects estimate, and therefore arguably a safer choice for estimating the population effect size (Carter et al., 2019, 2015). These differences in the baseline bias of the methods likely contributed to the differences in the extent of the effect size adjustment across the methods and the four models. We focused on ME and RMSE when evaluating the performance of each study. Other metrics, like coverage probability or Type I error rate, may be of interest to some researchers seeking to select a well performing adjustment method. The factors impacting method performance explored in this study were also limited. Each of the four scenarios determined which adjustment method should be used based on two assumed severity levels of publication bias (moderate vs high), and two assumed population effect sizes ($\delta = 0.2$ or $\delta = 0.5$). Although we argued that these levels of the two parameters are the most common in psychological research, researchers may wish to examine potential bias in their estimates when different values of these parameters are assumed. Similarly, level of questionable research practices (QRPs) assumed for our models was 0. The presence of QRPs (such as selective inclusion/exclusion of outliers, or selective use of covariates) can impact the performance of the estimation methods (Carter et al., 2019) as they further bias the sample by producing spurious effects, and there are compelling arguments to believe that some level of QRPs is common for meta-analyses (John et al., 2012). It is worthwhile considering the impact of QRPs on the performance when selecting a publication bias adjustment method. Likewise, it is worth bearing in mind that bias can enter meta-analyses in a multitude of ways on top of literature censoring and purposeful QRPs. The process of meta-analysing studies requires the researchers to make an abundance of analytic decisions, some of which will inevitably be biased by the researchers' beliefs about

the best analytic practice and may lead to errors made in good faith. Adequately addressing publication bias is therefore only one of the many challenges the researchers need to be prepared to confront when conducting a meta-analysis.

Summary and Conclusion

The aim of this study was to estimate the range of change in effect sizes when different publication bias adjustment methods are applied to real meta-analytic datasets matching the conditions necessary for the optimal performance of these estimators. The attenuation in the estimates was relatively small and usually not sufficient to alter the interpretation of the original results. This should not be taken as an indication of the lack of publication bias, as the studies of the prevalence of statistically significant results in the published literature consistently show that the level of publication bias in the field is high. By disregarding the problem, the meta-analysts may miss important moderators of the effects. There are a few steps the researchers can take to mitigate the bias in their own analyses, like thoroughly searching for unpublished studies, applying the adjustment methods only when sample sizes are adequate, and exploring the range of plausible values of the effects under study after being adjusted with appropriate methods. Nevertheless, publication bias remains a problem best addressed as a problem of the field and prevented where possible. While recent developments have been successful in addressing some of the wider structural problems in psychology contributing to biases in the publication process, sustained collective effort is still needed to improve the credibility of psychology as a science.

References

Allen, C., & Mehler, D. M. A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biology*, *17*(5), e3000246. https://doi.org/10.1371/journal.pbio.3000246

Avery, R. A. T., Kamar, M., Sladekova, M., & Field, A. P. (2020). How hast he last decade influenced meta-analytical reporting and practice quality. *Unpublished Manucript*, *https://osf.io/rovhd*.

Bartos, F., Maier, M., & Wagenmake, E.-J. (2020). *Adjusting for Publication Bias in JASP âE" Selection Models and Robust Bayesian Meta-Analysis* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/75bqn

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 1088–1101.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2011). *Introduction to Meta-Analysis*. John Wiley & Sons.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*(4), 796–815. https://doi.org/10.1037/xge0000083

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Chambers, C. (2019). *The seven deadly sins of psychology: A manifesto for reforming the*

culture of scientific practice. Princeton University Press.

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19*(2), 211–229. https://doi.org/10.1037/a0032968

Cohen, J. (1992). Things I have learned (so far). *Annual Convention of the American Psychological Association, 98th, Aug, 1990, Boston, MA, US; Presented at the Aforementioned Conference.*

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Cumming, G. (2013). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis* (1st ed.). Routledge. https://doi.org/10.4324/9780203807002

Duval, S. (2005). The trim and fill method. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, 127–144.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*(7109), 629–634.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS One*, *5*(4), e10068.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561.

Field, A. P. (2018). *Discovering statistics using IBM SPSS statistics. 5th Ed.* California: Sage Publication.

Field, A. P., Lester, K. J., Cartwright-Hatton, S., Harold, G. T., Shaw, D. S., Natsuaki, M. N., Ganiban, J. M., Reiss, D., Neiderhiser, J. M., & Leve, L. D. (2020). Maternal and paternal influences on childhood anxiety symptoms_ A genetically sensitive comparison. *Journal of Applied Developmental Psychology*, 12.

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy*, *98*, 19–38.

Gabry, J. (2018). *Shinystan: Interactive visual and numerical diagnostics and posterior analysis for bayesian models*. https://CRAN.R-project.org/package=shinystan

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1.

Gronau, Q. F., Erp, S. V., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*(1), 123–138. https://doi.org/10.1080/23743603.2017.1326760

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review*, *23*(1), 74–86. https://doi.org/10.3758/s13423-015-0868-6

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hunter, J. E., & Schmidt, F. L. (1996). Cumulative research knowledge and social policy formulation: The critical role of meta-analysis. *Psychology, Public Policy, and Law*, *2*(2), 324.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Ioannidis, J. P. A. (2016). The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses: Mass Production of Systematic Reviews and Meta-analyses. *The Milbank Quarterly*, *94*(3), 485–514. https://doi.org/10.1111/1468-0009.12210

Ioannidis, J. P. A. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis: Heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice*, *14*(5), 951–957. https://doi.org/10.1111/j.1365-2753.2008.00986.x

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*(3), 245–253.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Kim, N. Y., Bangdiwala, S. I., Thaler, K., & Gartlehner, G. (2014). SAMURAI: Sensitivity analysis of a meta-analysis with unpublished but registered analytical investigations (software). *Systematic Reviews*, *3*(1), 27.

Koller, M., & Stahel, W. A. (2011). Sharpening Wald-type inference in robust regression for small samples. *Computational Statistics & Data Analysis*, *55*(8), 2504–2515. https://doi.org/10.1016/j.csda.2011.02.014

Kontopantelis, E., & Reeves, D. (2012). Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A comparison between

DerSimonianLaird and restricted maximum likelihood: *Statistical Methods in Medical Research*. https://doi.org/10.1177/0962280211413451

Lau, J., Ioannidis, J. P. A., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *BMJ*, *333*(7568), 597–600. https://doi.org/10.1136/bmj.333.7568.597

Maier, M., Bartos, F., & Wagenmake, E.-J. (2020). *Robust Bayesian Meta-Analysis: Addressing Publication Bias with Model-Averaging* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/u4cns

Maki, A., Cohen, M. A., & Vandenbergh, M. P. (2018). Using Meta-Analysis in the Social Sciences to Improve Environmental Policy. In W. Leal Filho, R. W. Marans, & J. Callewaert (Eds.), *Handbook of Sustainability and Social Science Research* (pp. 27–43). Springer International Publishing. https://doi.org/10.1007/978-3-319-67122-2_2

McElrath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*(5), 730–749.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van't Veer, A. E., & Vazi, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., & Lakens, D. (2014). Registered Reports. *Social Psychology*, *45*(3), 137–141.

https://doi.org/10.1027/1864-9335/a000192

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, *8*(2), 157–159.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., Rushton, L., & Moreno, S. G. (2010). Assessing publication bias in meta-analyses in the presence of between-study heterogeneity: Assessing Publication Bias in Meta-analyses. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(3), 575–591. https://doi.org/10.1111/j.1467-985X.2009.00629.x

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638.

Sanchez-Meca, J., Marin-Martinez, F., & Chacon-Moscoso, S. (2003). Effect-Size Indices for Dichotomized Outcomes in Meta-Analysis. *Psychological Methods*, *8*(4), 448–467. https://doi.org/10.1037/1082-989X.8.4.448

Scheel, A. M., Schijen, M., & Lakens, D. (2020). *An excess of positive results: Comparing the standard Psychology literature with Registered Reports.* https://doi.org/10.31234/osf.io/p6e9c

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681.

Stanley, T. D. (2008). Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection*. *Oxford Bulletin of Economics and*

Statistics, *70*(1), 103–127. https://doi.org/10.1111/j.1468-0084.2007.00487.x

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, *8*(5), 581–591.

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *The American Statistician*, *49*(1).

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rucker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*(jul22 1), d4002–d4002. https://doi.org/10.1136/bmj.d4002

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*(11), 1119–1129.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How Many Studies Do You Need?: A Primer on Statistical Power for Meta-Analysis. *Journal of Educational and Behavioral Statistics*, *35*(2), 215–247. https://doi.org/10.3102/1076998609346961

Van Aert, R. C., & Van Assen, M. (2018). Correcting for publication bias in a meta-analysis with the p-uniform* method. *Manuscript Submitted for Publication Retrieved from: Https://Osfio/Preprints/Bitss/Zqjr92018*.

Van Aert, R. C., Wicherts, J. M., & Van Assen, M. A. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PloS One*,

14(4), e0215052.

Van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293.

Van Erp, S., Verhagen, J., Grasman, R. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 19902013. *Journal of Open Psychology Data*, *5*(1).

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2020). Rank-normalization, folding, and localization: An improved r-hat for assessing convergence of MCMC. *arXiv:1903.08008 [Stat]*. http://arxiv.org/abs/1903.08008

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods*, *10*(4), 428.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://www.jstatsoft.org/v36/i03/

Wilcox, R. R. (2016). *Introduction to Robust Estimation and Hypothesis Testing* (4 edition). Academic Press.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, *57*(1), 173–181.

Appendix A:

Directions and magnitudes of the effect size changes for meta-analyses across models and adjustment methods.

Model 1: moderate publication bias, $\delta = 0.2$

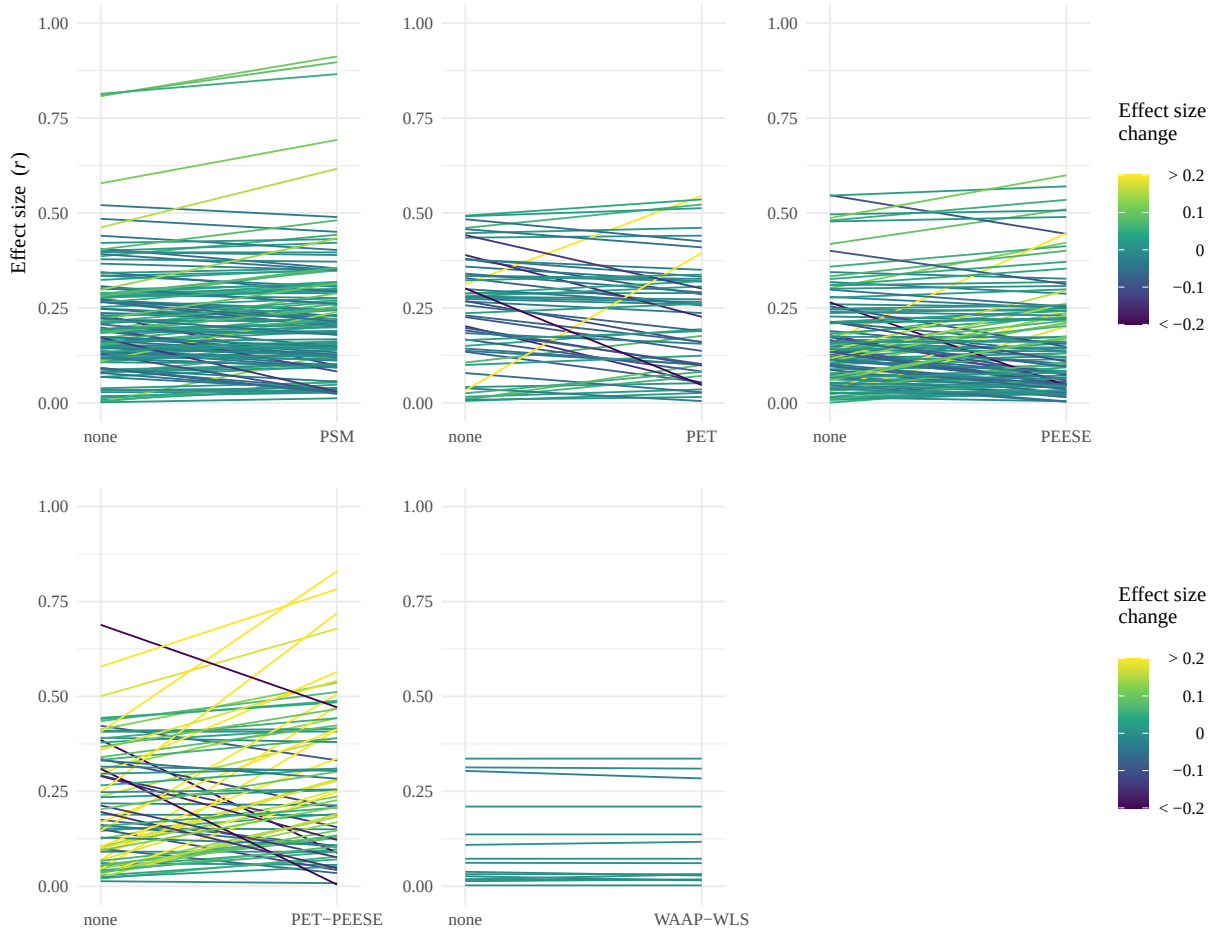

Figure 8

Effect size changes for meta-analyses in Model 1. Each line represents an individual meta-analysis. Slope and shade of the lines represent direction and magnitude of the change. Note. *3PSM and 4PSM were merged into a single factor level. Trim-and-fill was not evaluated in Model 1.*

Model 2: high publication bias, $\delta = 0.2$

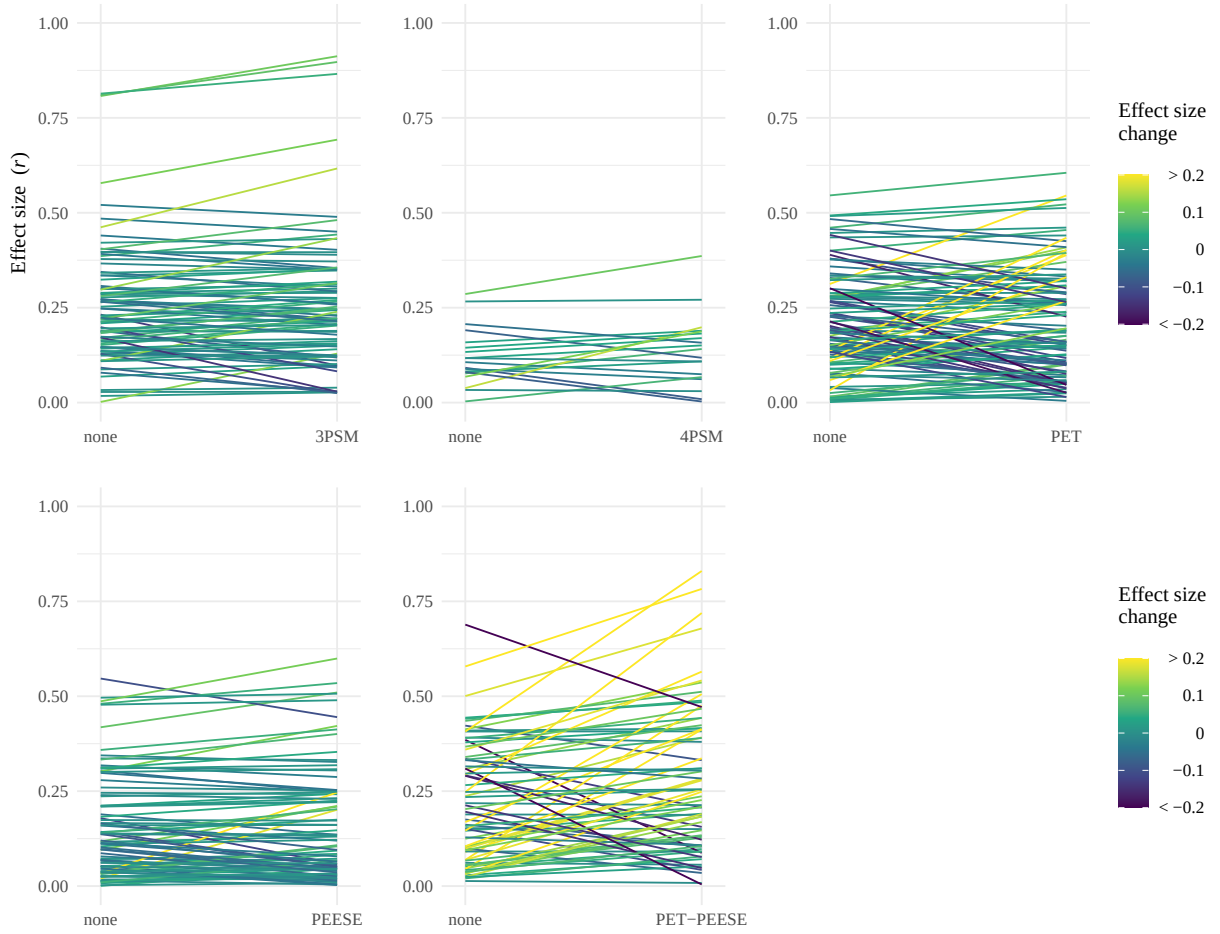

Figure 9

Effect size changes for meta-analyses in Model 2. Each line represents an individual meta-analysis. Slope and shade of the lines represent direction and magnitude of the change. Note. *Trim-and-fill and WAAP-WLS were not evaluated in this model.*

Model 3: moderate publication bias, $\delta = 0.5$

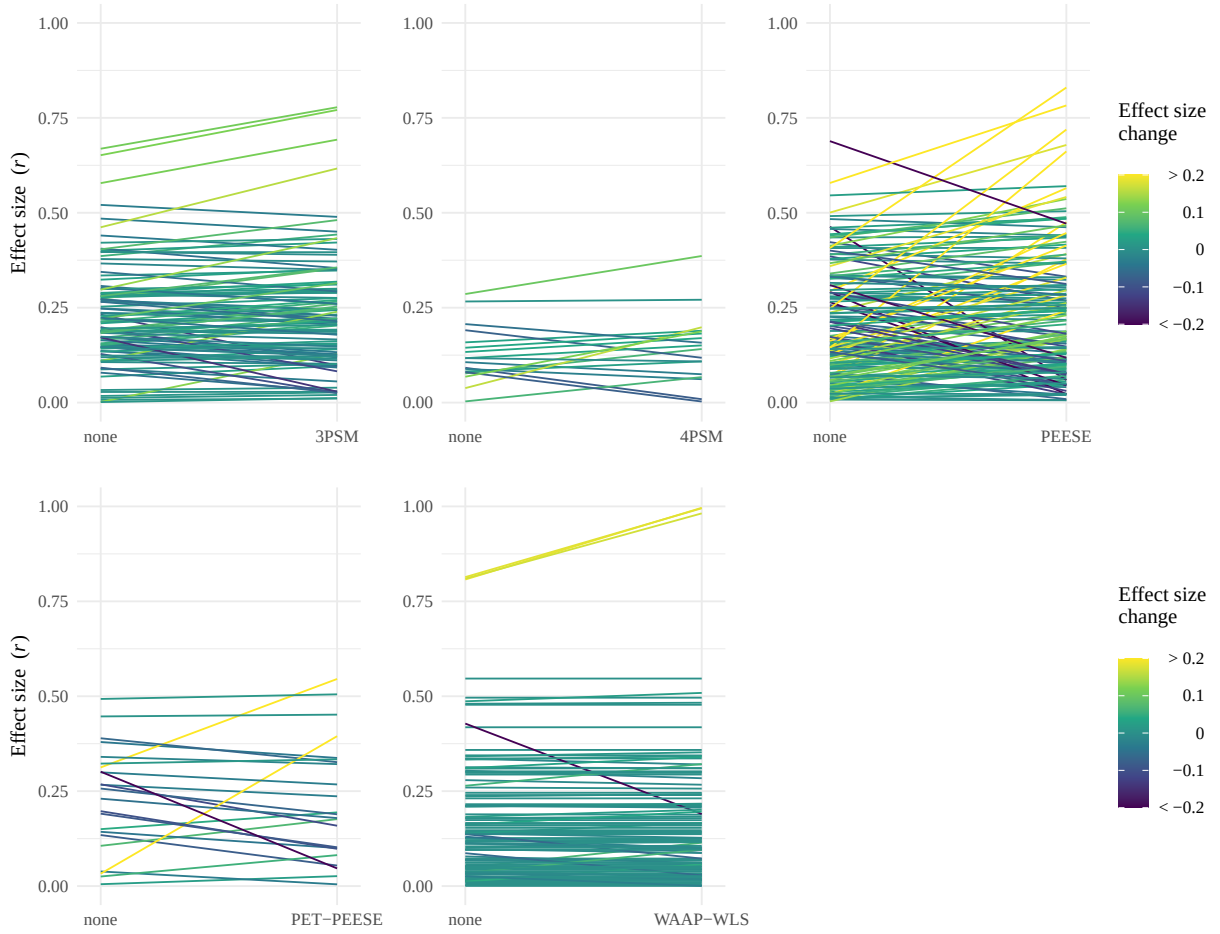

Figure 10

Effect size changes for meta-analyses in Model 3. Each line represents an individual meta-analysis. Slope and shade of the lines represent direction and magnitude of the change. Note. *PET and Trim-and-fill were not evaluated in this model.*

Model 4: high publication bias, $\delta = 0.5$

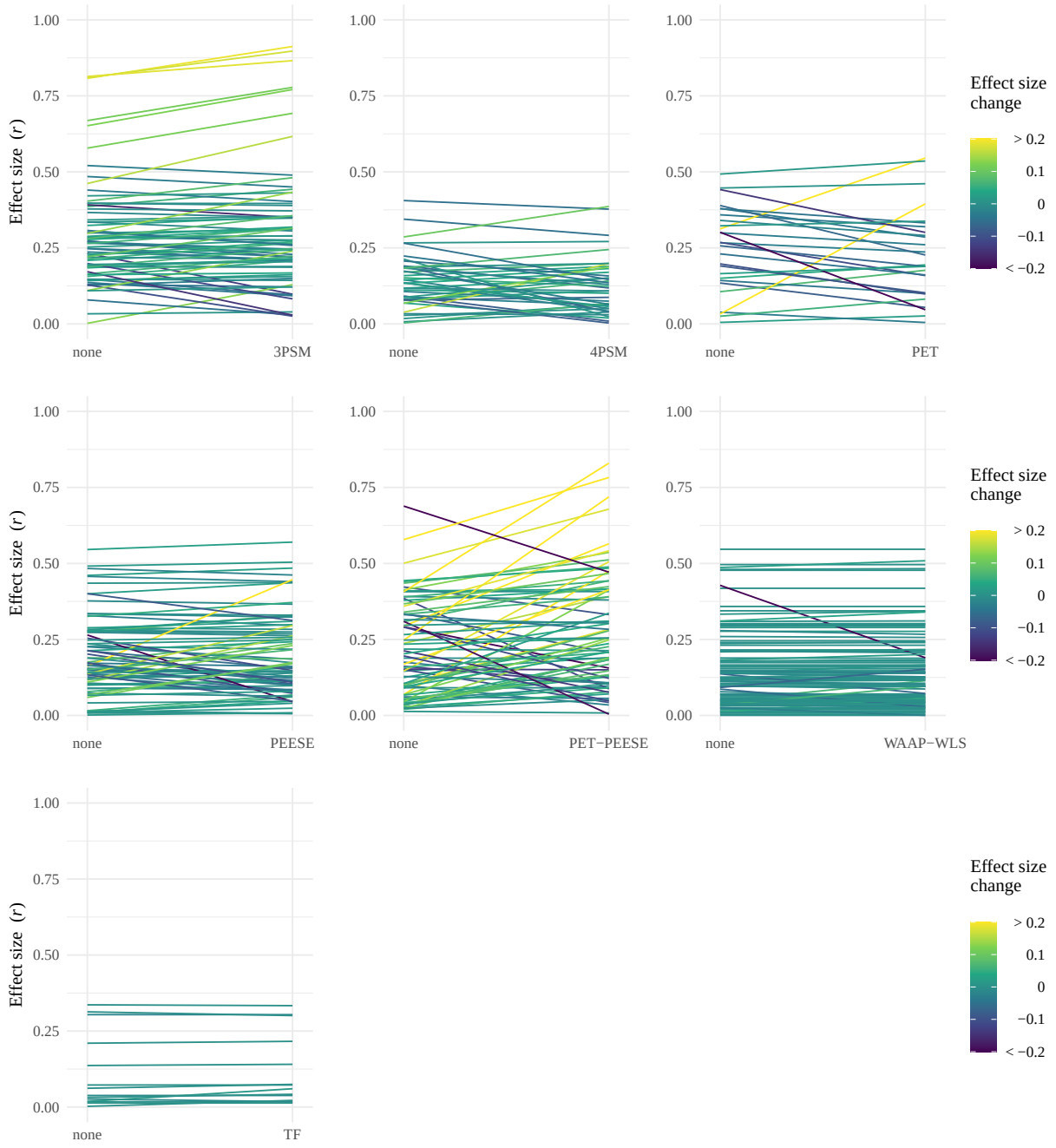

Figure 11

Effect size changes for meta-analyses in Model 4. Each line represents an individual meta-analysis. Slope and shade of the lines represent direction and magnitude of the change.